Recherche d'information et contexte

Gilles Hubert

MÉMOIRE

en vue de l'obtention de l'

Habilitation à diriger des recherches

délivrée par

l'Université Toulouse 3 – Paul Sabatier

Discipline INFORMATIQUE

présentée par

GILLES HUBERT

Recherche d'information et contexte

---

soutenue le 9 décembre 2010 devant la commission d'examen :

JURY

Catherine BERRUT	Professeure, Université Grenoble 1	*rapporteure*
Claude CHRISMENT	Professeur, Université Toulouse 3	*examinateur*
Mounia LALMAS	Professeure, Université de Glasgow (Écosse)	*rapporteure*
Thérèse LIBOUREL	Professeure, Université Montpellier 2	*examinatrice*
Josiane MOTHE	Professeure, IUFM Midi-Pyrénées	*directrice de recherche*
Jean-Marie PINON	Professeur, INSA de Lyon	*rapporteur*

École doctorale : Mathématiques, Informatique et Télécommunications de Toulouse

Unité de recherche : Institut de Recherche en Informatique de Toulouse – IRIT UMR 5505 CNRS

Équipe d'accueil : Systèmes d'Informations Généralisés

Mémoire composé avec LaTeX 2ε.

À ma mère qui nous a quittés soudainement
l'an passé et qui accompagne mes pensées.

Remerciements

Mes remerciements vont tout d'abord à Josiane Mothe, Professeure à l'IUFM de Toulouse, amie de longue date, pour toutes les collaborations partagées durant la dizaine d'années écoulée ; j'ai beaucoup appris sur la direction des recherches à son contact.

Je tiens ensuite à remercier Claude Chrisment, Professeur à l'université Paul Sabatier de Toulouse, pout m'avoir accueilli au sein de l'équipe il y a de nombreuses années et pour les conseils précieux qu'il m'a donnés tout au long de ces années.

Je remercie vivement Catherine Berrut, Professeure à l'université Grenoble I, Mounia Lalmas, Professeure à l'université de Glasgow et Jean-Marie Pinon, Professeur à l'INSA de Lyon, pour l'honneur qu'il m'ont fait en acceptant d'être rapporteurs de mon travail, à la lumière de leur expertise et pour leur participation au jury.

Je voudrais remercier également Thérèse Libourel, Professeure à l'université Montpellier II, pour me faire bénéficier de son expérience en acceptant d'examiner mon travail et de participer au jury.

Mes remerciements vont également à mes amis et collègues Franck Ravat, Olivier Teste, Max Chevalier, et Guillaume Cabanac avec qui j'ai pu partager collaborations fructueuses, moments de joie et également pour certains, moments de bricolage. J'adresse un merci spécial à Guillaume (même si par moment j'en suis moins sûr !) pour m'avoir initié à cet outil très particulier qu'est LaTeX. J'adresse également un grand merci à mes amis extérieurs à mon travail, Steff et Valérie, pour leur amitié et les bons moments passés en leur compagnie. Je tiens à remercier très sincèrement tous mes amis pour leur soutien dans les moments difficiles que j'ai traversés l'an passé ; leur amitié est précieuse et j'espère ne jamais la perdre.

Je n'oublie pas les palois Christian Sallaberry et Damien Palacio, pour nos collaborations fructueuses et nos grands moments de convivialité, les toulousains Ronan Tournier et Damien Dudognon, la lyonnaise Cécile Favre et l'italienne Antonella Dattolo pour leur aide et les moments que nous partageons, professionnels ou conviviaux.

Je voudrais exprimer toute ma gratitude aux membres de l'équipe SIG, passés, présents, et plus spécialement les permanents, pour leur soutien, leurs conseils et leur gentillesse. J'adresse également mes remerciements aux étudiants qui ont contribués à mes recherches.

Je voudrais dire merci à Agathe, Jean-Pierre, Arnaude, Jean-Claude et les autres collègues de l'IRIT qui, par leur bonne humeur et leur aide au quotidien, contribuent à nous permettre de mener nos activités de recherche. Je souhaite également remercier mes collègues tarbais qui ont contribué et contribuent encore, par leur sympathie et leur bonne humeur, à rendre mes déplacements sur Tarbes moins pesants.

Que ceux que je n'ai pas cité, et qui à leur manière m'ont apporté leur aide, m'excusent et soient remerciés même maladroitement.

Enfin, je ne peux clore ses remerciements sans faire une place spéciale à ma famille : ma mère, trop tôt disparue, pour tout ce qu'elle m'a apporté, mon père pour son soutien constant, ma sœur pour se soucier régulièrement de ma vie, mon épouse et nos trois filles pour tout le bonheur qu'elles m'apportent et qui m'aide à avancer.

Table des matières

Introduction générale

Contexte général des recherches

Les travaux présentés dans ce mémoire s'inscrivent dans le domaine de la recherche d'information (RI). Le domaine de la RI, apparu dans les années 1950-1960, n'a cessé d'évoluer jusqu'à aujourd'hui. Dès le début, l'évaluation des performances a été une préoccupation centrale par exemple avec le projet Crandfield (Cleverdon, 1967) qui a posé les bases de l'expérimentation en RI. Les performances mesurées se sont concentrées sur la qualité sémantique des réponses c'est-à-dire l'évaluation de la distance entre la pertinence système (implémentée par le système) et la pertinence utilisateur (jugée par l'utilisateur) (Denos, 1997). Elles se sont traduites par l'introduction des notions de rappel (nombre de documents pertinents pour l'utilisateur restitués par le système par rapport au nombre total de documents jugés pertinents) et de précision (nombre de documents pertinents pour l'utilisateur restitués par le système par rapport au nombre de documents pertinents pour le système).

La série de conférences TREC débutée en 1992 (Harman, 1992) a fortement contribué au développement de la RI en offrant des collections de test réalistes et des méthodes d'évaluations. Le schéma d'évaluation des performances mis en place a fortement conditionné le schéma d'évolution de la RI. Les recherches en RI ont été tout d'abord orientées système. Elles se sont d'abord concentrées sur la représentation des documents et des requêtes et leur mise en correspondance. Cela a conduit à la définition de modèles théoriques de RI comme le modèle vectoriel (Salton, 1971) et le modèle probabiliste (Robertson, 1977). Les propositions sont restées très longtemps axées sur l'appariement pour évaluer la correspondance entre les requêtes et les documents ainsi que sur l'indexation des documents et de requêtes pour obtenir une représentation qui supporte leur mise en correspondance.

L'objectif initialement visé a été de proposer un modèle de RI qui possède un comportement global le plus efficace possible. La RI s'est longtemps basée sur des hypothèses simplificatrices notamment en considérant un type unique d'interrogation et en appliquant le même traitement à chaque interrogation. Le contexte dans lequel s'effectue la recherche a été ignoré (Jones et Brown, 2004) à commencer par l'utilisateur (Kelly, 2009).

Le champ d'application de la RI n'a cessé de s'étendre notamment grâce à l'essor d'internet. Le volume d'information toujours plus important (l'espace numérique représente près de 500 exaoctets et doublera tous les 18 mois selon une étude économique (Gantz et Reinsel, 2009)) combiné à une utilisation de systèmes de recherche d'information (SRI) qui s'est démocratisée (près de 450 %

d'augmentation du nombre d'utilisateurs d'internet entre 2000 et 2010 [1]) ont conduit à une diversité des situations. Cet essor a rendu plus difficile l'identification des informations correspondant à chaque besoin exprimé par un utilisateur, marquant ainsi les limites des approches de RI existantes.

Face à ce constat, des propositions ont émergé, visant à faire évoluer la RI en rapprochant l'utilisateur du système. Dès les années 1970, un premier type d'approche propose de réinjecter la pertinence utilisateur dans le processus de RI et reformuler la requête initiale (Rocchio, 1971). Dans les années 1980 apparaissent les approches qui utilisent la notion de profil utilisateur de l'utilisateur pour reformuler la requête (Korfhage et Chavarria Garza, 1982). Le profil représente les centres d'intérêt à long terme de l'utilisateur. La notion de contexte en RI est introduite au cours des années 1990 au travers de modèles théoriques comme celui proposé par Saracevic (1997). L'exploitation de ces modèles théoriques a cependant été délaissée. La notion de contexte a été largement utilisée, correspondant à des aspects très divers (Finkelstein *et al.*, 2001). Elle englobe généralement celle de profil utilisateur, ajoutant des aspects liés à la tâche et à l'information manipulée.

Dans le but de fédérer les travaux et proposer des SRI offrant plus de précision en réponse au besoin de l'utilisateur, le domaine de la RI contextuelle a récemment émergé (Allan *et al.*, 2003). L'objectif est de différencier les recherches au niveau des modèles de RI en intégrant des éléments de contexte susceptibles d'avoir une influence sur les performances du SRI. La notion de contexte est vaste et se réfère à toute connaissance liée à la recherche de l'utilisateur interrogeant un SRI. Un contexte rassemble plusieurs aspects comme :

- l'utilisateur,
- l'information,
- l'environnement,
- le système.

Dans chacun de ces aspects différents éléments peuvent être considérés tels que :

- les caractéristiques de l'utilisateur (par exemple, ses connaissances),
- les intentions de l'utilisateur lors de la recherche (par exemple, la tâche justifiant sa recherche),
- le domaine de l'information,
- la structure de l'information (documents non structurés, semi-structurés ou structurés),
- les caractéristiques de l'environnement (par exemple, le matériel utilisé),
- les caractéristiques du système (par exemple, l'interface de visualisation).

Prendre en compte les éléments du contexte implique d'identifier et de modéliser ses éléments ainsi que de les intégrer dans les modèles et processus de RI. Mes travaux de recherche sont orientés vers l'exploitation de différents éléments de contexte suivant différentes stratégies. Ils s'inscrivent par conséquent dans le domaine de la RI contextuelle.

1. `http://www.internetworldstats.com/stats.htm`

Orientations de mes travaux de recherche

Mes recherches s'organisent suivant quatre orientations. Les trois premières orientations visent à intégrer des éléments de contexte tels que le domaine de l'information, la structure des documents et l'utilisateur. La quatrième orientation étudie l'adaptation du système en fonction du contexte.

Première orientation : Prendre en compte le domaine de l'information

Considérer le domaine de l'information dans le processus de RI est vu comme une réponse au problème d'ambigüité du langage naturel que ce soit au niveau de l'expression du besoin d'information ou de l'évaluation de la correspondance entre documents et besoins. Différents modèles ont été proposés pour décrire un domaine : les vocabulaires contrôlés, les taxonomies ou hiérarchies de concepts, les thésaurus, ou plus récemment les ontologies.

Les travaux de la littérature se sont principalement intéressés à la reformulation de requêtes pour la recherche *ad hoc* (c'est-à-dire lorsque le SRI restitue un ensemble de documents en réponse à une requête exprimée par l'utilisateur) et à la classification de documents pour le filtrage d'information (c'est-à-dire l'extraction d'informations répondant à un besoin au sein d'un flot de données).

D'autres problématiques moins explorées pour la recherche *ad hoc*, comme l'utilisation de la représentation du domaine comme moyen d'accès à l'information et de description des documents, m'ont conduit à mener des travaux pour y répondre. Mes recherches ont visé, d'une part, à proposer un modèle qui supporte l'ensemble des aspects d'une recherche d'information suivant des hiérarchies de concepts qui traduisent différentes facettes du domaine. Les problématiques traitées concernent la représentation multi-concept des documents suivant une ou plusieurs hiérarchies ainsi que par concepts, en texte libre ou en combinant ceux-ci. Mes recherches ont consisté ensuite à proposer des solutions au problème de l'évolution du corpus et de l'ontologie décrivant le domaine au travers de la proposition d'un modèle d'indexation sémantique.

Les travaux menés se situent dans un cadre monolingue (c'est-à-dire lorsque les documents et la représentation du domaine utilisent la même langue) de collections de documents non structurés homogènes du point de vue thématique. Ils répondent à des problématiques soulignées dans le cadre de différents projets (cf. Annexe, section 1, projets IRAIA, eStage et DynamO).

Seconde orientation : Prendre en compte la structure de l'information

La RI s'est concentrée majoritairement sur leur contenu textuel des documents. La structure logique des documents (par exemple en sections et paragraphes) est longtemps restée implicite et par conséquent peu exploité dans le processus de RI. Des travaux ont cependant tenté de manipuler la structure des documents (Salton *et al.*, 1993). L'apport des langages de balisages pour la description de documents, avec son premier représentant SGML (norme ISO 8879 :1986), a rendu explicites les informations concernant la structure des documents. De plus en plus de collections de documents décrits par un langage de balisage sont accessibles, notamment avec l'adoption du langage XML comme standard industriel pour l'échange de données (Fuselier et Chidlovskii, 2006). Considérer la structure des documents en plus du contenu textuel à introduit de nouvelles possibilités que ce soit au niveau de l'expression des besoins qu'au niveau des résultats restitués par les SRI. Les problématiques soulevées sont liées à la multi-granularité des éléments documen-

taires manipulées et à leur imbrication.

La recherche d'information dans des collections XML (RI XML) véhicule de telles problématiques. La proposition du cadre pluriannuel INEX pour l'évaluation des SRI XML a encouragé la proposition de solutions aux problématiques de la RI XML. Pour répondre à de telles problématiques en lien avec les projets Ws-Talk et QUEST que nous avons menés (cf. Annexe, section 1), mes recherches se sont attachées à faire des propositions en matière de modèle de RI au niveau de l'indexation des documents et des requêtes ainsi qu'au niveau de l'appariement entre requêtes et éléments XML.

Troisième orientation : Prise en compte de l'utilisateur

La plupart des SRI appliquent un traitement des requêtes de manière identique quel que soit l'utilisateur à l'origine de la requête. Les propositions de SRI ont eu principalement pour objectif de faire progresser leurs performances en améliorant les processus d'indexation et d'appariement, indépendamment des utilisateurs. Considérer l'utilisateur a cependant été perçu comme une possibilité d'amélioration. Les propositions de la littérature ont fait leur apparition au début des années 1980 (Korfhage et Chavarria Garza, 1982). Elles se sont principalement intéressées à la représentation de l'utilisateur sous la forme d'un profil utilisateur notamment du point de vue de ses centres d'intérêt et à la réinjection des informations dans les nouvelles requêtes.

Une problématique importante, peu traitée dans la littérature, reste l'évolution du profil de l'utilisateur face à leurs recherches afin qu'il reste représentatif de celui-ci, qu'il ne devienne pas omnipotent, et plus généralement qu'il n'en arrive pas à dégrader les performances du SRI plutôt que les améliorer.

Pour répondre à cette problématique, nos recherches se sont orientées vers une approche de gestion de profils d'interrogation et de leur évolution en étudiant notamment les conséquences de la modification des profils sur les performances du SRI. Les travaux menés, liés à des problématiques identifiées au sein d'un projet (cf. Annexe, section 1, projet IRAIA), se situent dans le cadre d'une recherche *ad hoc* sur des documents monolingues, non structurés, et axés sur un domaine spécifique du point de vue thématique.

Mes recherches se sont ensuite intéressées, dans un cadre plus général à la réutilisation d'expériences de recherche réalisées par les utilisateurs. Une proposition consiste à exploiter les requêtes précédemment posées pour aider un utilisateur à définir de manière incrémentale d'une nouvelle requête.

Quatrième orientation : Adapter le système

Les propositions qui prennent en compte des éléments de contexte telles que des profils d'utilisateurs ou des représentations de domaine s'inscrivent majoritairement dans un type d'approche qui consiste à étendre les traitements de RI. De nouvelles données liées au contexte sont traitées mais les traitements restent identiques pour toutes les recherches effectuées. Un autre type d'approche cherche cependant à modifier les traitements appliqués. C'est notamment le cas des travaux sur l'apprentissage automatique qui cherchent à identifier la fonction de score conduisant à la meilleure précision pour une collection donnée. Cependant de nombreuses questions soulevées dans le projet ACRIC (cf. Annexe, 1) restent en suspens liées à la variabilité des SRI telles que comment identifier les critères qui influencent les performances de SRI et comment agir sur les paramètres des SRI en fonction de ces critères. D'autres questions concernent l'identification

de l'interface de visualisation adaptée à une recherche donnée ou encore le choix du SRI le plus efficace pour une recherche.

C'est pour répondre à ces questions que j'ai mené des travaux pour étudier les critères relatifs aux requêtes qui peuvent expliquer la variabilité d'un SRI lors d'une recherche *ad hoc* et proposer une approche d'adaptation d'un SRI en fonction de préférences liées à l'utilisateur. Mes recherches se sont également intéressées à la sélection du SRI le plus efficace pour une recherche donnée ainsi qu'à la sélection de l'interface de restitution appropriée à une recherche.

Organisation du mémoire

Ce mémoire s'articule de la manière suivante :

- le chapitre 1 présente mes recherches pour prendre en compte le domaine lors de la RI après avoir introduit les travaux proposés dans la littérature. Il présente également les projets IRAIA, eStage et DynamO (cf. Annexe, 1) ainsi que les expérimentations menées sur des collections de référence et qui ont permis de valider mes contributions.
- le chapitre 2 est consacré à la prise en compte de la structure des documents. Il dresse un panorama des propositions existantes et détaille mes contributions sur cet aspect ainsi que les projets Ws-Talk et QUEST (cf. Annexe, section 1) qui ont permis de les valider. Il présente également les résultats des expérimentations obtenus dans le cadre de mes participations au programme d'évaluation pluriannuel INEX.
- le chapitre 3 s'intéresse à la prise en compte de l'utilisateur dans la RI. Il présente dans un premier temps les travaux de la littérature. Il expose ensuite nos propositions en matière de gestion de profils d'interrogation dont le projet IRAIA (cf. Annexe, section 1) a constitué un cadre applicatif. Il présente également mes contributions en matière de capitalisation d'expérience de recherche et le prototype *Query Explorer* développé dans ce cadre.
- le chapitre 4 traite de l'adaptation du système en fonction du contexte. Après avoir introduit les travaux proposés dans la littérature, il détaille nos propositions en matière d'adaptation de système. Ces propositions ont été validées dans le cadre du projet ACRIC (cf. Annexe, section 1) ou mises en œuvre au travers du prototype de plateforme d'évaluation d'interface de visualisation de résultat de RI, baptisé *VSE* et présenté dans ce chapitre.
- L'annexe (p. 155) synthétise les projets, les encadrements et les publications dans lesquels j'ai été impliqués.

1 Considérer le domaine lors de la recherche d'information

1 Introduction

Le volume d'information disponible électroniquement est toujours plus important et trouver des documents pertinents est une tâche de plus en plus délicate. L'ambigüité du langage naturel contribue à cette difficulté tant en termes d'expression du besoin d'information que de l'évaluation de la correspondance entre documents et besoins. Les multiples sens des termes et leurs multiples utilisations dans des domaines très variés participent également à rendre la tâche de recherche délicate. Considérer le domaine dans le processus de recherche d'information peut être une solution au problème d'ambigüité du langage naturel.

Ce chapitre est organisé de la manière suivante. La section 2 présente un panorama des travaux de la littérature proposés pour répondre aux problématiques soulevées et introduit nos propositions. La section 3 détaille nos propositions relatives à la RI guidée par les connaissances sur le domaine. La section 4 dresse le bilan de nos contributions et des perspectives envisagées.

2 Problématique et travaux de la littérature

Différents modèles ont été proposés pour décrire un domaine : les vocabulaires contrôlés, les taxonomies ou hiérarchies de concepts, les thésaurus, ou encore les ontologies en plein essor depuis quelques années. Ces modèles diffèrent en termes de degré de formalisation et du potentiel d'inférence des vocabulaires contrôlés jusqu'aux ontologies, plus riches, en passant par les hiérarchies de concepts puis les thésaurus :

- un vocabulaire contrôlé est un ensemble de termes,
- une taxonomie hiérarchise les termes suivant la relation « est-un »,
- un thésaurus relie les termes par différentes relations sémantiques (hiérarchiques, associatives, ou d'équivalence),
- enfin, une ontologie définit le sens des concepts par les termes qui les désignent et par une représentation structurée ou formelle de leur rôle dans le domaine (Aussenac-Gilles et Mothe, 2004).

Pour exploiter les connaissances sur le domaine, les propositions en RI se sont basées sur une représentation du domaine suivant l'un des modèles. Les connaissances sont ensuite intégrées

à des niveaux divers : dans l'expression de requête, dans la représentation de document, et dans l'appariement entre requête et document. Différentes exploitations de connaissances sur le domaine se distinguent :

- la reformulation de l'expression du besoin exprimé par l'utilisateur suivant les connaissances du domaine,
- l'interrogation par navigation suivant les connaissances du domaine,
- l'indexation sémantique des documents.

Les sections suivantes présentent les travaux de la littérature relatifs à ces trois types d'exploitation de connaissances sur le domaine.

2.1 Exploitation de connaissances sur le domaine pour l'expression du besoin

Parmi les approches utilisant les connaissances du domaine pour la reformulation de requêtes, certaines utilisent les ontologies comme moyen de désambiguïsation des termes de la requête (Guha *et al.*, 2003; Nagypál, 2005). Par exemple, Nagypál (2005) construit un ensemble de requêtes suivant différentes heuristiques pour ensuite fusionner les résultats obtenus. D'autres travaux substituent les termes de la requête par des concepts de l'ontologie suivant une analyse morphologique et sémantique (Aufaure *et al.*, 2007) ou par des vecteurs de termes associés aux concepts de l'ontologie (Tomassen *et al.*, 2006). Kwon *et al.* (1994) pondèrent les relations entre termes d'un thésaurus en fonction de la collection de documents pour ensuite étendre la requête avec des termes liés aux termes de la requête au-dessus d'un certain poids. Bodner et Song (1996) proposent un mécanisme d'expansion de requête par des termes liés dans un réseau sémantique représentant une base de connaissances spécifique d'un domaine. Bouchard et Nie (2006) se basent sur la construction d'un modèle de langue spécifique au domaine d'intérêt de l'utilisateur pour étendre sa requête initiale. Revuri *et al.* (2007) définissent différents cas d'expansion de requêtes avec différents éléments d'une ontologie de domaine tels que les concepts, les propriétés et les instances. Bhogal *et al.* (2007) présentent un panorama d'approches d'expansion de requête utilisant des connaissances de domaine.

Dans le domaine médical, Büttcher *et al.* (2004) reformulent les requêtes en ajoutant des variantes lexicales de termes médicaux et des synonymes d'acronymes ou symboles médicaux à l'aide de bases spécifiques (ex. AcroMed). Friberg (2007) s'appuie sur le thésaurus MeSH pour étendre une requête par des mots simples intervenant dans la construction de certains mots composés de langue suédoise. Plusieurs stratégies d'expansion de requêtes à partir du méta-thésaurus UMLS sont étudiées dans (Wollersheim et Rahayu, 2005), afin de mesurer leur efficacité ainsi que les variables ayant une influence. UMLS est également utilisé pour substituer certains termes d'une requête par des synonymes « préférés » (Plovnick et Zeng, 2004) ou pour étendre la requête avec des concepts plus spécifiques (Le *et al.*, 2007). Díaz-Galiano *et al.* (2009) utilisent l'ontologie médicale MeSH pour étendre les requêtes utilisateurs avec des termes médicaux.

Tuominen *et al.* (2009) utilisent certaines relations d'ontologie générale, d'ontologie de domaines spécifiques et d'ontologie spatio-temporelle pour enrichir une requête avec certains concepts reliés. Fu *et al.* (2004) se focalisent sur l'enrichissement de requête en exploitant les relations spatiales (ex. « au nord de », « proche de ») d'ontologies géographiques.

Dans le domaine juridique, Schweighofer et Geist (2007) introduisent un principe d'expan-

sion de requête dans un modèle de recherche booléen à partir d'une ontologie lexicale. Dans le domaine du développement durable, différentes mini-ontologies liées à plusieurs disciplines scientifiques sont fédérées pour créer une connaissance commune afin de contrôler la création de métadonnées et améliorer l'expressivité des requêtes des utilisateurs (Barde *et al.*, 2005b,a).

Squirrel (Duke *et al.*, 2007) est un outil de recherche et de navigation qui vise à combiner recherche en langage libre et recherche sémantique. L'utilisateur débute sa recherche par une requête en langage libre pour laquelle il obtient des documents associés à des propriétés liées à une ontologie. L'utilisateur peut ensuite raffiner sa requête en sélectionnant des éléments de l'ontologie parmi les propriétés des documents retournés. Cependant les deux modes de construction de la requête sont séparés et toujours utilisés séquentiellement dans cet ordre ; Duke *et al.* (2007) soulignent un besoin pour une meilleure intégration entre manipulation de l'ontologie et recherche en langage libre.

Ainsi, ces différentes propositions exploitent les connaissances sur le domaine après une formulation de requête initiale en langage libre. Un deuxième type d'exploitation consiste à utiliser directement les représentations des connaissances sur le domaine dont nous présentons différentes propositions dans la section suivante.

2.2 Exploitation de connaissances sur le domaine pour l'interrogation par navigation

L'exploitation de connaissances sur le domaine pour l'interrogation par navigation est proposée par exemple par les annuaires des moteurs de recherche tels que Yahoo ![1] initialement et Google [2] plus récemment. Ils proposent d'accéder aux documents à partir de catégories organisées hiérarchiquement. L'expression du besoin reste limitée au choix d'une catégorie et l'accès aux documents s'effectue majoritairement par navigation.

Des portails web sont d'autres exemples d'offre de navigation reposant sur une représentation du domaine. OntoWeb [3] est un portail d'échange d'informations basé sur une ontologie pour la gestion des connaissances et le commerce électronique. La recherche s'effectue en naviguant dans l'ontologie et les instances associées aux concepts. Une recherche « traditionnelle » *ad hoc* par mots-clés peut être réalisée mais indépendamment de l'ontologie.

Le portail web Medline [4] propose de définir une requête en cliquant sur les termes issus du thésaurus MeSH (de la bibliothèque nationale de médecine des États-Unis, « United States National Library of Medicine ») représenté sous forme d'arbre. Les documents indexés au préalable manuellement suivant ce même thésaurus MeSH et qui possèdent les termes sélectionnés sont ensuite restitués à l'utilisateur.

Dans le même esprit, Cat-a-Cone (Hearst et Karadi, 1997) est une interface 3D de recherche d'information suivant des hiérarchies de concepts. Les documents sont au préalable associés à différentes catégories de hiérarchies. Les utilisateurs définissent leurs requêtes en naviguant dans les hiérarchies représentées sous forme d'arbres 3D et en sélectionnant des catégories. En visualisant un document, la représentation 3D des hiérarchies montre également l'espace des concepts

1. `http://dir.yahoo.com/`
2. `http://directory.google.com/`
3. `http://www.ontoweb.org/`
4. `http://medline.cos.com/`

correspondant au document. Une recherche par mots-clés est possible pour chercher des catégories ou les documents suivant leur titre ou leur résumé.

Stuckenschmidt *et al.* (2004) proposent une exploration de publications basée sur un thésaurus de sciences naturelles et des métadonnées des documents telles que les auteurs. L'exploration s'effectue par l'intermédiaire d'une représentation graphique qui repose sur le regroupement des documents suivant les termes choisis par l'utilisateur et suivant le partage de ces termes entre les documents. La recherche est initiée par le choix d'un unique terme.

Avec l'interface CIRI (« Concept-based Information Retrieval Interface ») (Airio *et al.*, 2004), l'utilisateur peut sélectionner des concepts à partir d'ontologies disponibles affichées sous forme d'arbres. Il peut également supprimer certains termes issus des concepts des ontologies et ajouter ses propres termes pour ensuite soumettre la requête à des moteurs de recherche disponibles et des bases de données. La recherche mise en œuvre est une recherche traditionnelle *ad hoc* sur le contenu textuel des documents, ceux-ci n'étant pas indexés suivant les ontologies.

Shiri *et al.* (2007) développent également une interface visuelle basée sur un thésaurus pour la recherche d'information. À partir d'un terme, les concepts du thésaurus liées sont affichées et l'utilisateur peut sélectionner des concepts pour construire progressivement sa requête et obtenir ensuite les documents correspondant aux concepts choisis.

L'indexation sémantique suivant une ontologie ou une hiérarchie de concepts est bien souvent réalisée manuellement par des experts du domaine pour garantir la qualité de l'indexation (Hearst et Karadi, 1997; Handschuh *et al.*, 2002; Vallet *et al.*, 2005). Cependant, ce principe est lourd et coûteux notamment en temps. Les approches présentées dans la section suivante proposent donc d'automatiser ce procédé.

2.3 Exploitation de connaissances sur le domaine pour l'indexation sémantique

La catégorisation automatique de texte constitue un premier type d'approche pour associer des documents à des catégories prédéfinies. Ces catégories peuvent être organisées en hiérarchies. La catégorisation consiste à associer un document à une ou plusieurs catégories comme l'indexation sémantique suivant des hiérarchies de concepts.

De nombreux travaux basés sur un apprentissage ont été proposés. Lorsqu'un ensemble d'apprentissage est disponible, c'est-à-dire un ensemble de documents pré-catégorisés, celui-ci est utilisé pour optimiser l'outil de catégorisation. Lorsque l'apprentissage est terminé, de nouveaux documents peuvent alors être catégorisés, en fonction de leur adéquation avec les catégories apprises. Les algorithmes de type SVM (Support Vector Machine) (Vapnik, 1995) sont utilisés pour catégoriser des documents, montrant leur efficacité aussi bien dans le cas de catégorisation plane (Joachims, 1998), que dans le cas de catégorisation hiérarchique (Dumais et Chen, 2000). Néanmoins, dans le cas de grandes collections, la complexité de ces algorithmes devient trop importante et nécessite l'introduction de méthodes d'optimisation liées au problème de programmation quadratique (Chin, 1999). Dans le cas de gros volumes de documents ou de cadre applicatifs spécifiques, la simplicité et la robustesse du modèle de Rocchio peuvent s'avérer intéressantes (Vinot et Yvon, 2003; Vinot, 2004). Le modèle de Rocchio a été utilisé initialement dans le cadre de la reformulation de requête par réinjection de pertinence (Rocchio, 1971). De nombreuses adaptations de ce modèle pour la catégorisation ont été proposées (Lewis *et al.*, 1996; Joachims, 1997).

D'autres modèles issus de la statistique ou de l'apprentissage automatique sont utilisés pour la catégorisation (Sebastiani, 2002) comme les k plus proches voisins (k-NN) (Yang, 1994), les modèles Bayésiens (Koller et Sahami, 1997; Lewis et Ringuette, 1994), les algorithmes de « boost » (Schapire *et al.*, 1998), ou encore les réseaux de neurones (Schütze *et al.*, 1995).

Ces algorithmes ont été principalement utilisés pour la catégorisation mono-label c'est-à-dire pour associer un document à une catégorie unique. La catégorisation multi-label a longtemps été considérée comme une combinaison de catégorisations binaires (choix d'une catégorie unique entre deux catégories). Des propositions plus récentes s'appuyant principalement sur les techniques précédentes se focalisent plus spécialement sur la catégorisation multi-label basées par exemple, sur une méthode de régression de matrice (Sandu Popa *et al.*, 2007), ou sur un classifieur linéaire (Chang *et al.*, 2008).

Les approches basées sur des ontologies s'appuient généralement sur la recherche, dans les documents, des labels ou termes désignant les concepts ou instances (Desmontils et Jacquin, 2002; Pouliquen *et al.*, 2002; Kiryakov *et al.*, 2004). En effet, un concept peut être associé à plusieurs labels correspondant à autant de variantes lexicales qui désignent le concept. Zargayouna et Salotti (2004) prennent également en compte la structure et les fréquences d'apparition et de cooccurrence des termes pour indexer des documents XML suivant une ontologie de domaine. Dans le domaine médical, Le *et al.* (2007) étend l'index des documents par des concepts plus généraux issus du méta-thésaurus UMLS.

D'autres travaux proposent des environnements pour annoter les documents tout en peuplant l'ontologie (Vargas-Vera *et al.*, 2002; Amardeilh, 2006).

2.4 Orientation de nos contributions

Dans le cadre de la prise en compte du domaine dans la RI, les propositions relatives à la reformulation de requêtes constitue une part importante des travaux de la littérature (Kwon *et al.*, 1994; Aufaure *et al.*, 2007; Bhogal *et al.*, 2007). Dans ces approches, l'utilisateur ne manipule pas directement la connaissance sur le domaine. Quelques travaux ont proposé d'utiliser la représentation du domaine comme support pour exprimer les besoins d'information (Hearst et Karadi, 1997; Airio *et al.*, 2004; Shiri *et al.*, 2007). Cependant, ces travaux s'appuient sur une indexation manuelle des documents suivant les concepts de hiérarchies ou d'ontologies. Une réponse pour automatiser cette indexation a été faite par l'intermédiaire des méthodes de catégorisation (Joachims, 1998; Sebastiani, 2002) mais se sont concentrés sur l'association d'un document à une catégorie unique. Une limite de ces méthodes est la nécessité d'un apprentissage préalable. Un autre type d'approche dans lequel nous nous situons consiste à rechercher les labels des concepts dans les documents (Desmontils et Jacquin, 2002; Kiryakov *et al.*, 2004).

Nos travaux ont été orientés vers l'utilisation de la représentation du domaine pour exprimer son besoin d'information. Nous avons toutefois considéré un intérêt à pouvoir exprimer son besoin à la fois à l'aide de concepts issus du domaine et de texte libre. De plus, nous avons cherché à proposer des solutions pour automatiser tous les processus sous-jacents à une RI basée sur une représentation du domaine sous forme de hiérarchie de concepts ou d'ontologie. Nos travaux se sont donc orientés vers :

– la proposition d'un modèle de recherche qui supporte une recherche d'information suivant

des hiérarchies de concepts. Ce modèle permet notamment la catégorisation multi-label de documents et la recherche à partir d'un ensemble de catégories. Il permet également une recherche *ad hoc* en langage libre et la combinaison de recherche mêlant langage libre et catégories prédéfinies.
– la proposition d'un modèle d'indexation sémantique dynamique suivant une ontologie c'est-à-dire tenant compte de l'évolution du corpus et de l'ontologie.

3 Contributions pour la RI utilisant les connaissances sur le domaine

Les connaissances d'un domaine sont représentées sous diverses formes comme par exemple des thésaurus, des hiérarchies de concepts ou des ontologies. En RI, ces représentations sont majoritairement utilisées pour reformuler une requête exprimée par l'utilisateur en langage libre. Il s'agit en général d'éviter l'utilisation de termes ambigus en les remplaçant par des termes issus du domaine. La requête reformulée est ensuite utilisée pour effectuer une recherche *ad hoc*.

Notre proposition vise à utiliser directement la représentation de la connaissance pour définir le besoin d'information et représenter les documents. Elle consiste en la définition d'un modèle de RI :

– adaptable à une catégorisation automatique pour représenter les documents à l'aide de concepts organisés en hiérarchies,
– adaptable à une recherche *ad hoc* suivant des concepts,
– adaptable à une recherche *ad hoc* en texte libre.

Nous montrons également que la combinaison des modes de recherche en texte libre et par concepts est également plus efficace que l'utilisation séparée de ces modes (Hubert et Mothe, 2009). Des expérimentations sur des collections largement utilisées montrent l'intérêt de nos propositions (Augé *et al.*, 2001, 2003; Hubert et Mothe, 2009). Ces principes ont également été mis en œuvre dans le cadre des projets européens IRAIA [5] (Mothe *et al.*, 2002a,b) et eStage [6] (Englmeier et Mothe, 2002).

Une évolution de ces travaux concerne l'indexation et la recherche suivant des ontologies. Dans ce contexte, une problématique à laquelle nous tentons d'apporter des solutions concerne la gestion de l'évolution de l'indexation sémantique suite à l'évolution des documents et des ontologies. Le projet ANR DynamO [7] constitue un cadre applicatif à ces travaux.

3.1 Recherche suivant des hiérarchies de concepts

Le principe de recherche par concepts repose sur l'existence au préalable de concepts et leurs associations aux documents d'une collection. L'utilisateur sélectionne une catégorie pour ensuite accéder aux documents qui y sont associés. Notre approche consiste plutôt à utiliser les catégories comme moyen donné à l'utilisateur pour décrire son besoin en information. En effet, générale-

5. IRAIA : Getting Orientation in Complex Information Spaces as an Emergent Behaviour of Autonomous Information Agents, European project IST-1999-1062, 5th Framework Programme.

6. eStage : A New Stage for the Cultural Heritage in European Puppetry, European project IST-2000-28314, 5th Framework Programme.

7. Dynamo : DYNAMic Ontology for Information Retrieval, Projet ANR-07-TLOG-004.

ment une seule hiérarchie de concepts est utilisée pour accéder aux documents. Notre approche repose sur l'utilisation de plusieurs hiérarchies de concepts comme autant de facettes pour décrire les documents. Ce principe rend plus coûteux le recours généralement fait à une catégorisation manuelle des documents c'est-à-dire l'association manuelle des documents aux catégories par des experts. C'est pourquoi nous basons également notre approche sur une catégorisation automatique des documents c'est-à-dire l'association automatique des documents aux concepts de différentes hiérarchies.

3.1.1 Utilisation de la catégorisation pour l'accès aux documents

Dès lors que les documents sont au préalable associés à des hiérarchies que concepts, ces hiérarchies peuvent alors servir de moyen d'accès à ces documents. Les hiérarchies de concepts sont considérées comme des points de vue sur les documents qui peuvent être utilisées par l'utilisateur pour décrire son besoin en information. Pour cela, l'utilisateur navigue dans les différentes hiérarchies et choisit les catégories qui correspondent à son besoin. Une hiérarchie peut par exemple représenter une vue sur un domaine particulier. Cette notion est à rapprocher de celle de langage contrôlé et d'ontologie. Un même document peut être accessible par différents concepts de la même hiérarchie. Il peut également être accessible par des concepts appartenant à différentes hiérarchies de concepts (par exemple, différentes vues sur un domaine).

3.1.2 Définition d'un modèle générique de recherche d'information

L'objectif a été de mettre en place un environnement utilisable dans le contexte de recherche d'information suivant des hiérarchies de concepts, mais également de pouvoir l'utiliser dans d'autres contextes de RI comme la recherche en texte libre ou la recherche dans des collections spécifiques. À cet effet, nous avons défini un modèle générique de RI, présenté dans cette section et dont nous détaillons les applications à la catégorisation de documents et à la recherche de documents dans les sections suivantes.

Pour définir notre modèle générique de RI nous nous sommes basés sur l'architecture usuelle de SRI. Notre modèle est ainsi basé sur deux notions : la notion d'unité d'information (UI) et celle de besoin d'information (BI). Le principe d'un SRI est d'évaluer l'adéquation entre UI et BI. Nous avons choisi de distinguer ces deux notions en considérant que leurs rôles sont différents. Même si ces deux notions partagent certaines caractéristiques, celles-ci peuvent être exploitées différemment. Il peut exister également des caractéristiques spécifiques à chaque notion.

Notre approche se décompose suivant les différentes étapes :

- indexation des UI,
- indexation des BI,
- mesure de la correspondance entre UI et BI à l'aide d'une fonction de similarité,
- classement des UI suivant le score obtenu.

L'indexation des UI et des BI consiste à extraire les éléments caractéristiques notamment du point de vue du contenu textuel. La fonction de similarité consiste à évaluer les correspondances entre les éléments caractéristiques des UI et des BI. Les sections suivantes présentent, d'une part, l'indexation des UI et BI, et la mesure de correspondance d'autre part.

3.1.2.1 Indexation des unités d'information et des besoins d'information

Nous avons choisi de baser notre approche sur une représentation vectorielle (Salton *et al.*, 1975) des UI et des BI. Le modèle vectoriel a été largement adopté dans les SRI. Notre choix s'explique par la simplicité du modèle et son extensibilité.

Une unité d'information est définie comme l'élément de granularité minimale qui peut être recherché. Elle correspond habituellement à un document ou à une partie de document. Une unité d'information est définie par un identifiant et un ensemble de concepts (par exemple, des termes) comme suit :

$UI_i = \left(l_i, \{(c_1, w_1), \ldots, (c_j, w_j)\}\right)$

où l_i est l'identifiant de l'unité d'information, c_j est un concept décrivant l'unité d'information UI_i, et w_j est le poids donné au concept c_j pour UI_i (par exemple, le nombre d'occurrences de c_j).

Un besoin d'information est l'expression de critères que l'utilisateur souhaite voir apparaître dans les éléments du résultat. Les critères n'ont pas nécessairement la même priorité. Un BI est en général défini par un ensemble de concepts recherchés. Suivant les contextes, d'autres critères s'ajoutent à la définition d'un BI comme par exemple des indications de préférence associées aux concepts (par exemple, concepts souhaités ou concepts non souhaités).

La définition d'un besoin d'information est ainsi la suivante :

$BI_k = \{(c_m, w_m, cp_m), \ldots, (c_n, w_n, cp_n,)\}$

où c_m est un concept décrivant le besoin d'information BI_k, w_m est le poids attribué au concept c_m pour BI_k (par exemple, la fréquence), cp_k est la préférence associée au concept c_k pour le BI.

Par exemple, une indexation textuelle à partir les termes suivra un processus habituel d'extraction des termes représentatifs d'un document ou d'une requête en langage libre. L'extraction de terme inclut la suppression de mots vides et optionnellement des traitements comme la racinisation ou la lemmatisation. Pour les BI, des traitements supplémentaires comme l'extraction des préférences associées aux concepts recherchés sont mis en œuvre.

3.1.2.2 Appariement unité d'information - besoin d'information

De nombreux SRI basés sur le modèle vectoriel utilisent le cosinus comme mesure de similarité entre vecteurs. Comme le souligne Singhal (2001), le produit scalaire est souvent utilisé comme mesure de similarité alternative au cosinus qui offre plus de possibilités d'adaptation. Pour cette raison également, notre appariement est basé sur une fonction de score dérivée du produit cartésien. Celle-ci est fondée sur la somme des contributions des concepts partagés par BI et UI. Ce principe peut également être rapproché d'un principe de vote.

La fonction de score repose sur trois facteurs :

- l'importance d'un concept pour le BI.
- l'importance d'un concept pour l'UI recherchée.
- le recouvrement global entre BI et UI.

Elle est définie en les combinant de la manière suivante :

$$Score(BI_1, UI_1) = \left(\sum_{c \in C_1} imp_{BI}(c, BI_1) \cdot imp_{UI}(c, UI_1) \right) \cdot recouv(BI_1, UI_1) \quad (1.1)$$

où $imp_{BI}()$ dénote une fonction qui estime l'importance du concept c (appartenant à l'ensemble C_1 des concepts de BI_1 et UI_1) dans le besoin d'information BI_1. Cette importance est basée sur le poids w_j associé au concept lors de l'indexation du BI (cf. section3.1.2.1). imp_{UI} dénote une fonction qui estime l'importance du concept c dans l'unité d'information UI_1. Cette importance est basée sur le poids w_k associé au concept lors de l'indexation de l'UI. *recouv* dénote une fonction qui estime le recouvrement entre BI et UI. Par exemple, l'estimation peut se baser sur le nombre de concepts communs.

La fonction de score est une somme des contributions des concepts partagés par les besoins d'information et les unités d'information. Ce principe a pour but de favoriser la présence de concepts « importants » pour les unités d'information et pour les besoins d'information. Les définitions des fonctions imp_{BI} et imp_{UI} influencent le classement des unités d'information selon les termes qu'elles contiennent. La somme implique également que plus les unités contiennent des concepts « importants » plus leur score est élevé. Indirectement, les définitions des fonctions imp_{BI} et imp_{UI} vont également déterminer la manière de classer les unités d'information contenant peu de concepts « importants » par rapport à celles contenant beaucoup de concepts peu « importants ». Cependant, la mise en place d'une stratégie de classement de ces deux types d'UI l'un par rapport à l'autre est complexe. Pour cette raison, la fonction *recouv* est introduite, la fonction de score en estimant globalement le recouvrement entre UI et BI. Le but est d'offrir la possibilité de rééquilibrer le score suivant le nombre de concepts partagés. Ce principe permet d'adapter pour chaque recherche la manière d'évaluer la pertinence des UI contenant peu de termes « importants » par rapport aux UI contenant beaucoup de concepts moins « importants ».

Ce modèle est défini pour permettre différentes stratégies de recherche et être appliqué à différents types de corpus (par exemple, être documents structurés). Unités d'information et besoins d'information peuvent être traités de manières différentes en termes de contenus et de structure.

3.1.3 Application du modèle générique à la catégorisation automatique de documents

L'accès à des documents par l'intermédiaire de concepts suppose que ces documents soient préalablement associés aux concepts. Cette association peut être effectuée manuellement par des experts comme souvent le cas. Cependant l'association manuelle s'avère d'autant plus délicate lorsqu'il s'agit d'associer les documents à des catégories de plusieurs hiérarchies. Une alternative est d'effectuer cette association de manière automatique. C'est pourquoi nous avons proposé une méthode de catégorisation automatique (Augé *et al.*, 2001, 2003; Hubert et Mothe, 2009). La catégorisation automatique s'avère d'autant plus efficace qu'un apprentissage est possible. Nous avons donc étudié une méthode d'apprentissage adaptée à notre principe de catégorisation automatique.

3.1.3.1 Principe de catégorisation automatique

Le principe de la catégorisation automatique de documents peut être abordé suivant deux stratégies différentes :

– trouver les documents qui correspondent à un concept,
– trouver les concepts qui correspondent à un document.

Nous avons choisi la seconde manière pour définir notre approche compte tenu que nous considérons les concepts comme moyen de description des documents. De plus, ce type d'approche est plus adapté à un environnement où la collection de documents évolue constamment. La catégorisation peut simplement dans ce cas s'effectuer à chaque document ajouté.

Plutôt que d'étudier une solution spécifique au processus de catégorisation automatique, nous avons envisagé de mettre en place une approche réutilisable pour d'autres processus de recherche d'information. Nous avons opté pour un processus d'appariement usuel de recherche d'information dans le but de définir une base de SRI qui puisse être adapté à d'autres cas de figures. En effet, nous avons considéré que la catégorisation consiste à rechercher des UI que sont les concepts répondant à un BI représenté par un document.

Notre approche se décompose en 4 étapes :

– indexation des documents,
– indexation des concepts,
– évaluation de la correspondance entre document et concept à l'aide d'une fonction de score,
– association du document avec les concepts suivant le score obtenu en fonction d'une stratégie définie. La stratégie peut être d'associer tous les concepts obtenant un score dépassant un seuil fixé ou d'associer les concepts obtenant les scores les plus élevés.

L'indexation des documents suit un processus « usuel ». Le contenu textuel d'un document est examiné afin d'en extraire les termes représentatifs et leur nombre d'occurrences dans le document. L'extraction de termes met en œuvre notamment la suppression des mots vides en utilisant un anti-dictionnaire. Des traitements supplémentaires comme une racinisation ou une lemmatisation peuvent être également appliqués de manière optionnelle.

L'indexation des concepts suit le même processus d'extraction des termes représentatifs et de nombres d'occurrences dans le concept. L'indexation des catégories s'effectue de manière indépendante c'est-à-dire sans considérer l'organisation hiérarchique de spécialisation des catégories. Chaque catégorie est donc indexée suivant son propre contenu textuel. Éventuellement, l'organisation hiérarchique des concepts peut être considérée en étendant dans ce cas la représentation d'un concept par les termes des concepts ascendants qu'il spécialise ou descendants qu'il généralise.

3.1.3.2 Appariement document - concept pour la catégorisation

La fonction de score repose sur les trois facteurs introduits précédemment dans le modèle appliqué au contexte de catégorisation de documents :

– l'importance d'un terme pour le BI (fonction $imp_{BI}()$). Dans notre cadre de catégorisation de documents, le BI est représenté par un document. Ce facteur estime donc l'importance du terme pour le document,
– l'importance d'un terme pour l'UI recherchée (fonction $imp_{UI}()$). Dans notre cadre de catégorisation de documents, l'UI est représentée par un concept. Ce facteur évalue donc l'importance du terme pour le concept,
– le recouvrement global entre le BI et l'UI (fonction *recouv()*) c'est-à-dire entre le document

et le concept.

La fonction de score correspond à la définition ci-dessous :

$$Score(D,C) = \left(\sum_i f_{i,D} \cdot \frac{f_{i,C}}{cf_i} \right) \cdot \varphi^{\frac{N_{D,C}}{\min(N_D,N_C)}} \textit{ (Augé et al., 2001)} \tag{1.2}$$

où D est un document
C est un concept
$f_{i,D}$ Fréquence du terme t_i dans le document D
$f_{i,C}$ Fréquence du terme t_i dans le concept C
cf_i Nombre de concepts dans la hiérarchie qui contiennent le terme t_i
$N_{D,C}$ Nombre de termes communs au document D et à le concept C
N_D Nombre de termes distincts dans le document D
N_C Nombre de termes distincts dans le concept C
φ Réel positif $\geqslant 1$

L'importance du terme dans le document est estimée suivant sa fréquence dans le document (Salton, 1971). L'hypothèse est que plus le terme est fréquent dans le document plus le document traite de ce terme. Ce terme est d'autant plus représentatif du besoin d'information qu'il représente le document.

L'importance du terme pour un concept est définie par la fréquence du terme dans la catégorie divisée par le nombre de catégories contenant ce terme. L'hypothèse est que l'importance d'un terme dépend de son apparition dans le concept par rapport aux autres termes mais surtout dépend de son pouvoir discriminant entre les différents concepts. Cette dernière notion est la notion d'*idf* bien connue en recherche d'information (Spärck Jones, 1972). Exploiter le pouvoir discriminant des termes est particulièrement important dans le cadre de concepts hiérarchiquement organisés suivant un principe de généralisation/spécialisation. En effet, dans ce contexte, les termes d'un concept se retrouvent souvent dans les concepts qui le spécialisent.

Enfin le recouvrement global entre un document et une catégorie est estimé suivant la proportion de termes communs par rapport au maximum de termes communs possible appelé communément coefficient de recouvrement (« overlap coefficient ») (van Rijsbergen, 1975). L'hypothèse est bien sûr que plus il y a de termes en commun entre deux éléments plus le recouvrement entre les deux éléments est important. Ce coefficient est placé en exposant d'une constante pour accentuer les écarts entre catégories suivant le nombre de termes possédés. De plus, la valeur attribuée à la constante φ permet de faire varier l'influence de ce facteur sur le score final par rapport aux facteurs liés aux contributions individuelles des termes.

La notion de couverture complète la fonction de calcul de score afin de considérer dans le processus de recherche uniquement les éléments dont le contenu textuel renferme une part suffisante de concepts de la requête. La couverture est définie comme un seuil correspondant au pourcentage minimum de termes de la requête qui doivent apparaître dans le contenu textuel d'un élément pour que cet élément soit pris en compte dans le processus de recherche. La couverture vise à assurer que la recherche exploite seulement des UI où la requête est suffisamment représentée.

Le principe est d'exprimer de façon globale une exigence sur les critères exprimés dans la requête (par exemple, au moins 50 % des termes).

La couverture est intégrée à la fonction de score de la manière suivante :

$$Score(D,C) = \left(\sum_i f_{i,D} \cdot \frac{f_{i,C}}{cf_i} \right) \cdot c_{D,C} \cdot \varphi^{\frac{N_{D,C}}{\min(N_D,N_C)}} \text{ (Augé et al., 2001)} \quad (1.3)$$

tel que Si $\frac{N_{D,C}}{\min(N_D,N_C)} \geqslant CT$ Alors $c_{D,C} = 1,0$ Sinon $c_{D,C} = 0,0$

où CT Réel positif représentant le seuil tel que $0,0 \leqslant CT \leqslant 1,0$

$N_{D,C}$ Nombre de termes communs au document D et à la catégorie C

N_D Nombre de termes distincts dans le document D

N_C Nombre de termes distincts dans la catégorie C

Un intérêt de cette définition est de pouvoir faire varier la valeur de seuil à chaque application de la fonction de score et notamment pour chaque requête. Selon le nombre de concepts recherchés et les concepts indiqués l'utilisateur peut ainsi indiquer à chaque recherche s'il veut absolument que tous les concepts recherchés soient présents dans les éléments retrouvés ou seulement une certaine proportion.

3.1.3.3 Les contextes d'application des projets IRAIA et eStage

Cette section présente les deux projets qui ont constitué des cadres applicatifs de nos propositions.

Le projet européen IRAIA (Getting Orientation in Complex Information Spaces as an Emergent Behaviour of Autonomous Information Agents) visait à fournir aux utilisateurs un accès aisé à des informations telles que des documents textuels et des séries temporelles. Des documents appartenant à des langues différentes étaient manipulés. Les documents étaient semi-structurés (par exemple exprimés en HTML) ou non structurés (texte libre).

IRAIA a été développé dans un contexte de traitement d'informations économiques. Cependant, les méthodes développées peuvent être appliquées à différents domaines ; elles ont été évaluées dans cette optique.

Une des originalités du projet IRAIA a résidé dans le fait qu'il reposait sur des hiérarchies de concepts c'est-à-dire des arbres composés de nœuds correspondant à des concepts dénotés par une expression et de relations « est un » entre eux. Ces hiérarchies de concepts correspondant aux connaissances du domaine fournissent à l'utilisateur un espace de navigation sémantique. Pour interroger le système, l'utilisateur navigue dans les hiérarchies de concepts pour définir son besoin d'information. Celui-ci définit une requête avec une spécificité graduelle en descendant dans la hiérarchie. Les hiérarchies de concepts constituent donc le langage de communication c'est-à-dire le langage d'interrogation mais également le langage de représentation des documents quels que soient leur type et leur langue.

En conséquence tous les documents doivent être décrits suivant ces hiérarchies de concepts. Notre rôle a concerné l'indexation de la collection de documents dans un environnement multilingue dans le cadre du lot « analyse automatique de texte ». Dans cette partie du projet, nous avons apporté des solutions de catégorisation automatique telles que présentées précédemment.

Le projet IRAIA s'est appuyé sur trois hiérarchies de concepts traduisant trois vues sur le domaine économique :

- la hiérarchie *Branch* décrivant les branches économiques et industrielles,
- la hiérarchie *Country* décrivant les pays et regroupement économique de pays,
- la hiérarchie *Theme* décrivant les indicateurs économiques.

Les documents étaient rédigés en anglais, allemand ou français.

Le principe de navigation est le suivant :

- l'utilisateur sélectionne des concepts au sein des trois hiérarchies proposées dans le domaine économique puis soumettait sa requête,
- le système retourne comme résultat les documents décrits par les concepts sélectionnés,
- l'utilisateur peut visualiser le contenu d'un document du résultat et les concepts associés au document suivant chacune des trois hiérarchies,
- l'utilisateur peut compléter sa requête en sélectionnant des catégories lors de la visualisation d'un document.

Le projet eStage (A New Stage for the Cultural Heritage in European Puppetry) s'est appuyé sur l'infrastructure développée dans le projet IRAIA pour produire un nouveau service dédié au domaine culturel européen des marionnettes. eStage est un service d'information construit autour d'éléments liés à la marionnette tels que des pièces de théâtre, des descriptions des marionnettes, de la littérature comme des contes. L'idée a été de rendre disponible gratuitement une collection numérique sur les marionnettes qui vive de la libre contribution de sa communauté. Le principe d'eStage a été de construire un environnement qui permette d'organiser et de rechercher les contributions fournies par la communauté suivant des hiérarchies de concepts représentant le domaine de la marionnette. À l'instar du projet IRAIA, nous avons apporté, dans le projet eStage, des solutions de catégorisation automatique telles que présentées précédemment.

3.1.3.4 Expérimentations sur la catégorisation dans le cadre IRAIA

Le projet IRAIA a fourni des collections de documents du domaine économique et différentes hiérarchies de concepts traduisant différentes facettes de ce domaine. Pour évaluer notre catégorisation, nous avons extrait un ensemble de documents à catégoriser suivant trois hiérarchies de concepts représentant un domaine d'IRAIA. Ces hiérarchies ont été choisies car elles possèdent des caractéristiques différentes en termes de nombre de catégories, de nombre de termes par catégorie, et de profondeur d'arborescence. Le jeu de test utilisé pour ces expérimentations possède les caractéristiques suivantes :

Hiérarchie	Concepts	Termes	Termes par concept (moyenne)
Branch	423	1570	3,7
Country	23	25	1,1
Theme	30	96	3,2
Nombre de documents : 40			

Tableau 1.1 – Caractéristiques de la collection de test IRAIA

Une première phase a consisté à demander à un ensemble d'utilisateurs d'associer manuellement les documents HMTL (en langue anglaise traitant d'économie) aux catégories des trois hiérarchies. Ces juges pouvaient associer jusqu'à 10 catégories d'une hiérarchie avec une pertinence sur une échelle de un à cinq.

La catégorisation automatique réalisée par notre SRI a ensuite été comparée à la catégorisation manuelle préalablement établie. L'efficacité de la catégorisation automatique a été évaluée en utilisant les critères usuels de la recherche d'information : la précision et le rappel. Dans le cadre de la catégorisation, ces critères sont définis comme suit (Yang, 1999) :

$$rappel = \frac{\left|\{catégories\ pertinentes\ retrouvées\}\right|}{\left|\{catégories\ pertinentes\}\right|}, \quad précision = \frac{\left|\{catégories\ pertinentes\ retrouvées\}\right|}{\left|\{catégories\ retrouvées\}\right|}$$

Le programme trec_eval [8] a été utilisé pour calculer les valeurs de précision et de rappel. La précision est calculée suivant 11 points de rappel (0,0 ; 0,1 ; ... ; 1,0).

Le but de ces expérimentations a été avant tout de mesurer l'efficacité de notre approche suivant différents types de hiérarchies et valider notre contribution. La figure 1.1 illustre un extrait des résultats obtenus.

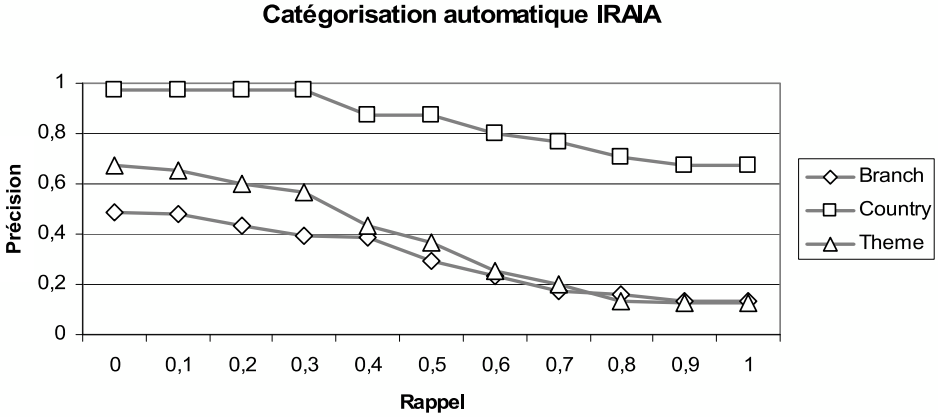


Figure 1.1 – Catégorisation automatique dans le cadre IRAIA (Augé *et al.*, 2001)

Ces expérimentations ont conduit à différentes observations notamment une efficacité de notre approche variable selon les caractéristiques de la hiérarchie. Pour des hiérarchies peu profondes, avec peu de concepts et des concepts ayant des descriptions totalement distinctes (comme pour la hiérarchie Country) notre approche est efficace mais ce cas est naturellement favorable. En présence de peu de catégories l'approche montre des résultats encourageants même si cette situation reste également plutôt favorable. En présence de nombreux concepts et une hiérarchie plus profonde, la sélection des « bons » concepts est plus difficile sans doute en raison des spécialisations plus nombreuses de concepts et par conséquent des concepts ayant des descriptions proches. Pour améliorer la catégorisation de ce type de hiérarchies et compte tenu de l'efficacité observée en présence de descriptions de concepts très distinctes, une solution explorée a été l'enrichissement de la représentation des concepts. Pour cela une phase d'apprentissage pourrait être utilisée pour ajouter des termes permettant de distinguer davantage les descriptions

8. http://trec.nist.gov

des concepts. La section suivante présente les principes de l'apprentissage que nous avons exploré et les expérimentations réalisées.

3.1.3.5 Expérimentations sur l'enrichissement de la description des concepts sur la collection Reuters-21578

Nous avons mené des expérimentations dans le but d'évaluer l'apport d'une phase d'apprentissage sur notre approche de catégorisation automatique. L'apprentissage consiste exploiter des cas d'utilisation d'un processus avérés réussis ou échoués dans le but d'améliorer les performances de ce processus. Dans le cadre de la catégorisation, l'apprentissage est un moyen d'enrichir la description d'un concept par de nouveaux termes. Il s'agit d'ajouter des termes liés aux termes initiaux en identifiant des liens de type « associé à ». Lorsque des documents déjà catégorisés sont jugés comme correctement catégorisés, une analyse du contenu textuel de ces documents peut fournir des termes qu'il serait intéressant d'ajouter aux descriptions des concepts afin d'améliorer la catégorisation de nouveaux textes. Cette analyse peut être basée sur des indicateurs statistiques comme les fréquences des termes dans les documents associés aux différents concepts. Notre approche a consisté à appliquer une fonction qui associe un poids à chaque terme candidat issus des documents catégorisés. Les termes avec les plus poids les plus élevés ont ensuite été ajoutés à la description d'un concept (Augé *et al.*, 2003).

Ces expérimentations ont été menées sur le corpus Reuters-21578 [9] largement utilisé pour évaluer les techniques de catégorisation et de classification (Lewis *et al.*, 1996; Joachims, 1998). La qualité du corpus a été améliorée par David Lewis en supprimant entre autres les documents dupliqués, les erreurs typographiques et les catégorisations erronées. Le corpus modifié nommé « ModLewis » est composé d'une collection d'apprentissage (13625 documents) et une collection de test (6188 documents) et 135 catégories. L'apprentissage est réalisé sur la collection correspondante et l'efficacité est évaluée sur la collection de test. Cependant cette collection ne permet pas de traiter des concepts organisés hiérarchiquement puisque les concepts de Reuters ne sont pas hiérarchisés. Pour ces expérimentations les critères d'évaluations usuels basés sur les notions de rappel et précision ont été appliqués (cf. section 3.1.3.4).

Ces expérimentations ont mis en œuvre notre méthode de catégorisation sans apprentissage (nommée CP) et après une phase d'apprentissage (nommée CPLearn50). Les résultats obtenus avec notre approche ont été comparés à une méthode « traditionnelle » de catégorisation (nommée RocchioCos) combinant la mesure cosinus pour l'association de concepts et le modèle de Rocchio (Rocchio, 1971) pour la phase d'apprentissage.

La figure 1.2 résume les résultats représentatifs des expérimentations menées sur le corpus Reuters-21578.

Ces résultats montrent qu'une phase d'apprentissage appropriée améliore considérablement les résultats en moyenne notre méthode de catégorisation (+ 97 % pour CPLearn50 par rapport à CP). De plus, notre approche avec apprentissage aboutit, en moyenne, à des résultats nettement supérieurs à une combinaison (Rocchio+Cosinus) largement utilisée (+ 36 % pour CPLearn50 par rapport à RocchioCos). Les résultats obtenus avec notre approche sont nettement améliorés avec un enrichissement de la description des concepts par 50 termes. En dessous de 50 termes ajoutés, les résultats sont améliorés mais dans une moindre mesure. Ajouter davantage de termes conti-

9. http://www.daviddlewis.com/resources/testcollections/reuters21578/

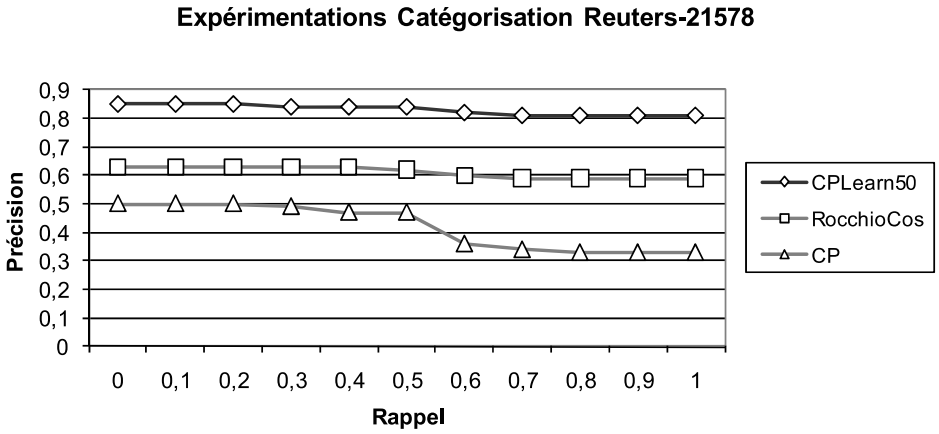


Figure 1.2 – Catégorisation automatique sur le corpus Reuters-21578 (Augé *et al.*, 2003)

nue à faire progresser les résultats mais dans une moindre mesure (par exemple, l'ajout de 150 termes améliore les résultats de + 1,6 % par rapport à l'ajout de 50 termes). Les catégories doivent donc être suffisamment enrichies pour aboutir à un processus de catégorisation efficace. Ajouter ensuite des termes supplémentaires ne conduit pas significativement à mieux décrire les catégories.

3.2 Combinaison des recherches par concepts et par mots-clés

Selon Fidel (1991), la recherche à partir de concepts et la recherche en texte libre sont deux modes complémentaires plutôt qu'opposés. Fugmann (2002) déclare également qu'un système combinant les deux types de recherche serait intéressant afin d'exploiter leurs forces et de compenser leurs faiblesses. Cependant, l'immense majorité des systèmes ne propose qu'un seul des deux modes plutôt que les deux, et principalement la recherche en texte libre. De plus, les systèmes existants qui incluent les deux modes proposent essentiellement de choisir un mode unique à chaque recherche. Combiner les deux modes permettrait à l'utilisateur de sélectionner les concepts qu'il identifie clairement comme correspondant à son besoin d'information et de compléter sa requête à l'aide de mots-clés pour lesquels il n'a pas identifié de concept. Inversement, une telle combinaison permettrait à l'utilisateur de préciser sa requête en texte libre avec quelques concepts sans avoir l'obligation de sélectionner un ensemble complet de concepts.

Partageant l'idée de complémentarité des deux modes de recherche, nous avons étudié le moyen d'adapter notre modèle générique de RI (cf. section 3.1.2) pour permettre la recherche suivant les deux modes et leur combinaison (Hubert et Mothe, 2009). La combinaison des modes n'est a priori destinée qu'à traiter des besoins d'information mélangeant langage libre et catégories. L'approche est basée sur des représentations appropriées des documents et des besoins d'informations et des définitions des fonctions de score pour l'appariement adaptées à chacun des deux modes.

Notre approche reprend l'idée poursuivie par les travaux relatifs à la combinaison de résultats de recherches (Fox et Shaw, 1994; Lee, 1997; Montague et Aslam, 2002) et auxquelles nous nous intéressons également (Hubert et Mothe, 2007b,a; Hubert *et al.*, 2007b) (cf. chapitre 3). Cette idée est que deux SRI peuvent être complémentaires et que dans ce cas combiner les résultats qu'ils

fournissent améliore la recherche.

Du point de vue de l'utilisateur la combinaison des modes peut être réalisée de différentes manières, par exemple :

- L'utilisateur commence sa recherche suivant les concepts. Il sélectionne tout d'abord des concepts correspondant à son besoin d'information à partir d'une représentation graphique des concepts suivant le principe proposé dans le système Cat-a-Cone (Hearst et Karadi, 1997) ou le système OntoExplo (Hernandez *et al.*, 2005). L'utilisateur complète ensuite l'expression de son besoin au travers d'un texte libre. Inversement, il peut débuter en langage libre l'expression de son besoin et le compléter en sélectionnant des concepts.
- L'utilisateur commence sa recherche en langage libre ou à l'aide de concepts. Des documents sont retournés à l'utilisateur et à partir de la visualisation des documents, l'utilisateur peut compléter sa requête en sélectionnant des concepts associés aux documents visualisés (Aussenac-Gilles et Mothe, 2004) ou en sélectionnant des termes issus des contenus des documents.

3.2.1 Application du modèle générique à la recherche ad hoc en texte libre

La recherche *ad hoc* consiste généralement à trouver un ensemble de documents correspondent à un besoin d'information exprimé en texte libre. Il s'agit dans ce cas d'adapter notre modèle générique (cf. section 3.1.2) en considérant les documents comme les unités d'information et les requêtes comme les besoins d'information.

Dans le contexte de documents non structurés et de requêtes en texte libre, le processus d'indexation inclut la suppression des mots vides et éventuellement un processus de racinisation ou de lemmatisation. Les poids des termes associés aux termes dans les index sont, par exemple, les nombres d'occurrences. Le processus d'indexation permet ainsi d'obtenir les représentations de type UI et BI adéquates.

En s'appuyant sur les valeurs statistiques communément utilisées en RI (Manning *et al.*, 2008, ch. 8), nous avons appliqué notre fonction de score générique à la recherche en texte libre en définissant les trois facteurs de la manière suivante :

- l'importance d'un terme pour le besoin d'information (dans ce contexte, une requête) est estimée par sa fréquence. L'hypothèse est que plus un terme apparaît dans la définition de la requête, plus il est représentatif du besoin,
- l'importance d'un terme pour l'unité d'information (dans ce contexte, un document) est estimée d'un part par sa fréquence et d'autre part par son pouvoir discriminant entre les unités d'information. L'hypothèse est d'une part que si un terme est fréquent dans le document (hors mots vides) il est d'autant plus représentatif du document. D'autre part, le principe est celui de la valeur statistique idf qui considère que plus un terme est rare dans les autres unités d'information plus il est important pour distinguer les UI dans lesquels il apparaît,
- enfin, le facteur estimant globalement le recouvrement entre UI et BI est basé sur le coefficient de recouvrement à l'instar de la définition pour la catégorisation (cf. section 3.1.3.2). Il est basé sur la proportion de termes communs par rapport au nombre de termes distincts dans le plus petit élément entre UI et BI en l'occurrence BI. Ce coefficient est élevé à la puis-

sance pour accentuer l'écart entre les UI en fonction du nombre de termes du BI qu'elles possèdent.

La fonction d'appariement est donc la suivante :

$$Score(BI, UI) = \left(\sum_i \frac{tf_{t_i,BI}}{tf_{\max,BI}} \cdot \frac{tf_{t_i,UI}}{tf_{\max,UI} \cdot e_{t_i,C}} \right) \cdot \varphi^{\frac{N_{BI,UI}}{\min(N_{BI},N_{UI})}} \text{ (Hubert et Mothe, 2009)} \quad (1.4)$$

Où *BI* est un besoin d'information et *UI* une unité d'information. $tf_{t_i,BI}$ dénote le nombre d'occurrences du terme t_i dans *BI* et $tf_{\max,BI}$ dénote le nombre maximum d'occurrences d'un terme dans *BI*. $tf_{t_i,UI}$ correspond au nombre d'occurrences de t_i dans l'unité d'information *UI* et $tf_{\max,UI}$ dénote le nombre maximum d'occurrences d'un terme dans *UI*. $e_{t_i,C}$ dénote le nombre d'UI du corpus *C* (i.e., l'ensemble des UI) qui contiennent le terme t_i. $N_{BI,UI}$ est le nombre de termes communs à *BI* et *UI* ; N_{BI} est le nombre de termes distincts dans *BI* et N_{UI} le nombre de termes distincts dans *UI*. φ est un réel supérieur ou égal à 1 qui permet de faire varier l'influence de ce facteur. Une valeur égale à 1 implique que ce facteur n'intervient pas dans le calcul de score.

Cette fonction d'appariement est basée principalement sur des pondérations très utilisées telles que tf.idf (Salton et Buckley, 1988). Cependant, nous ne reprenons pas la définition habituelle utilisant un logarithme en raison de l'introduction du facteur φ. Le logarithme dans la définition de l'*idf* permet de faire décroitre plus progressivement la valeur lorsque la présence du terme augmente dans le corpus. Ce principe, combiné à une somme implique une manière particulière d'ordonner les UI, en particulier celles contenant peu de termes rares par rapport à celles contenant de nombreux termes plus fréquents dans le corpus. Notre combinaison avec le facteur φ conserve le même principe tout en permettant un ordonnancement différent selon la valeur de φ comme illustré dans l'exemple 1.

Exemple 1. *Soit un corpus de 1500 documents, 6 termes définissant une requête (avec une occurrence de chaque terme), et trois documents contenant certains des 6 termes (avec une occurrence de chaque terme), comme suit :*

Terme	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*
Nombre de documents contenant le terme	*800*	*700*	*450*	*400*	*300*	*100*
D1		✓	✓	✓	✓	
D2	✓					✓
D3	✓	✓		✓	✓	

Le document D1 contient les termes {T2, T3, T4, T5}, D2 contient les termes {T1, T6}, et D3 contient {T1, T2, T4, T5}.

Différents classements peuvent être obtenus en appliquant l'équation 2 avec différentes valeurs de φ :

	Rang selon le score calculé		
Fonction de score	*D1*	*D2*	*D3*
Équation 2 avec la définition habituelle de l'idf (logarithme) au lieu de $\frac{1}{e_{t_i,C}}$ *et sans le facteur* φ *(*φ*=1)*	*1*	*3*	*2*
Équation 2 sans le facteur φ *(*φ*=1)*	*2*	*1*	*3*
Équation 2 avec φ*=2*	*1*	*2*	*3*
Équation 2 avec φ*=5*	*1*	*3*	*2*

La valeur de φ doit donc être fixée en fonction du cas souhaité.

3.2.2 Application du modèle générique à la recherche par concepts

La recherche par concepts consiste à fournir à l'utilisateur un ensemble de concepts prédéfinis d'une ou plusieurs hiérarchies de concepts. Dans ce contexte, l'utilisateur ne spécifie pas de requête en langage libre. Il navigue dans la structure pour accéder aux documents ou pour définir son besoin d'information. Notre approche est basée sur la représentation des documents par des concepts prédéfinis plutôt que des termes. L'ensemble de concepts représentant un document peut résulter d'une catégorisation manuelle par un expert du domaine ou par un processus de catégorisation automatique en appliquant par exemple la mise en œuvre de notre modèle (cf. section 3.1.3). La problématique est ensuite de fournir un moyen de retrouver les documents répondant à un ensemble de concepts définissant un besoin d'information.

D'une part, un index à partir des concepts prédéfinis peut être construit pour les documents. Chaque document est décrit par un ensemble de concepts éventuellement pondérés. D'autre part, des concepts sélectionnés par l'utilisateur représentent son besoin d'information. Dans ce cas, le processus de recherche consiste à calculer un score d'appariement entre :

- chaque unité d'information correspondant à un ensemble de concepts représentant un document
- et un besoin d'information correspondant à un ensemble de concepts sélectionnés par l'utilisateur.

Il est possible de définir l'appariement simplement par le nombre de concepts communs au besoin d'information et à l'unité d'information. Une autre solution que nous avons retenue est d'utiliser la fonction de score définie précédemment pour la recherche en texte libre (cf. section 3.2.1) en considérant que les concepts décrivant BI et UI sont décrits par des concepts au lieu de termes (Hubert et Mothe, 2009). Dans ce cas, la fréquence d'un concept dans un document ou une requête est en général égale à 1.

3.2.3 Combinaison des résultats des recherches par concepts et texte libre

La combinaison des recherches par concepts et texte libre n'a de sens que lorsque le besoin d'information est exprimé à la fois par des termes et par des concepts. Lorsque le besoin d'information n'est défini que par des termes l'appariement correspondant à la recherche en texte libre peut être appliqué seul ; lorsque le besoin d'information n'est défini que par des concepts l'appa-

riement correspondant à la recherche par concepts peut être appliqué seul.

La combinaison des modes de recherche peut supposer la définition d'une nouvelle fonction d'appariement mêlant facteurs liés aux termes et facteurs liés aux concepts.

Une autre solution, que nous avons retenue, s'appuie sur les techniques de combinaison de résultats de recherche (Fox et Shaw, 1994; Lee, 1997; Montague et Aslam, 2002). Ces techniques ont montré l'efficacité de combiner les résultats retournés par deux systèmes différents.

Nous appliquons la fonction CombSum proposée par Fox et Shaw (1994), largement utilisée, en raison de sa faible complexité et de son efficacité. La fonction CombSUM est appliquée sur les résultats obtenus séparément en calculant la somme des scores obtenus, d'une part, par la fonction de score entre UI et BI décrits par des termes (cf. section 3.2.1), et d'autre part, par la fonction de score entre UI et BI décrits par des concepts (cf. section 3.2.2).

Pour évaluer l'efficacité de la combinaison des modes de recherche par concepts et texte libre, nous avons réalisé des expérimentations sur le jeu de données d'apprentissage de la tâche Filtrage de la campagne d'évaluation TREC-9 (TREC-9 Filtering Track) [10]. Cette collection de test est composée du corpus OHSUMED-87, d'un ensemble de 63 requêtes et de leurs jugements de pertinence, et d'un ensemble de catégories MeSH. Le corpus OHSUMED-87 est un ensemble de 54710 documents MEDLINE de l'année 1987. Chaque document contient différents champs notamment ceux que nous avons utilisés pour nos expérimentations : le titre (« title »), le résumé (« abstract »), et les catégories associées (« assigned MeSH categories »). Chaque requête contient deux champs : le titre (« title ») et la description de l'information recherchée (« description »).

Nous avons choisi ce jeu de test parce que :

- il fournit, d'une part, un ensemble de requêtes pour la recherche *ad hoc* en langage libre avec les jugements de pertinence pour chaque requête,
- et d'autre part, les documents sont annotés avec des concepts ce qui offre la possibilité de réaliser une recherche par concepts.

Pour permettre la recherche par catégorie, chaque requête a été examinée par une personne suivant le titre et la description afin d'identifier un ensemble de concepts MeSH correspondant à chaque requête. Le but a été d'avoir, pour chaque requête définie pour une recherche *ad hoc* en texte libre, une requête équivalente utilisable pour une recherche par concepts. Les utilisateurs qui ont choisi manuellement les concepts étaient familiers des moteurs de recherche mais non familiers des concepts MeSH et du domaine médical du corpus. L'objectif n'était pas d'obtenir les concepts « idéaux » pour décrire chaque requête mais d'obtenir plutôt des cas « réels » de besoins d'information exprimés à la fois en texte libre et à l'aide de concepts.

Pour ces expérimentations, les évaluations se sont basées sur les critères habituels de rappel et précision définis par :

$$rappel = \frac{\left|\{documents\ pertinents\ retrouvés\}\right|}{\left|\{documents\ pertinents\}\right|}, \quad précision = \frac{\left|\{documents\ pertinents\ retrouvés\}\right|}{|\{documents\ retrouvés\}|}$$

Le programme trec_eval [11] a été utilisé pour calculer les valeurs de rappel et de précision. Pour

10. http://trec.nist.gov/data/filtering/README.t9.filtering
11. http://trec.nist.gov

chaque requête, la précision est calculée chaque fois qu'un nouveau document pertinent est restitué ; la somme des différentes précisions calculées est ensuite divisée par le nombre de documents pertinents pour la requête pour obtenir la précision moyenne AP (« Average Precision »). La MAP (« Mean Average Precision ») correspond à la moyenne des AP obtenues pour un ensemble de requêtes.

Une première étude a été menée pour comparer l'efficacité de la combinaison des deux modes par rapport à chaque mode seul. La figure 1.3 compare les trois processus de recherche suivant la précision moyenne pour toutes les requêtes :

- Ctexte désigne l'exécution de notre modèle suivant le mode de recherche en texte libre uniquement (cf. section 3.2.1, équation 1.4 avec $\varphi = 400$) ; la valeur de φ résulte d'autres expérimentations réalisées sur d'autres corpus en traitant uniquement le champ titre de la requête,
- Cconcept désigne l'exécution de notre modèle suivant le mode de recherche par concepts uniquement (cf. section 3.2.2, $\varphi = 400$),
- Ccomb désigne la combinaison des deux modes en appliquant la méthode CombSUM (Fox et Shaw, 1994) (cf. section 3.2.3) pour construire le résultat.

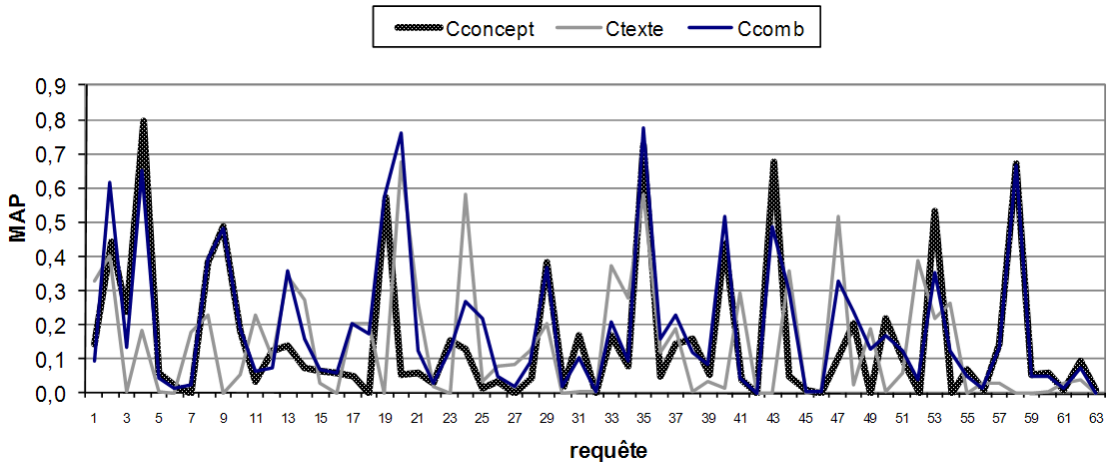


Figure 1.3 – MAP pour chaque mode séparé et les modes combinés (Hubert et Mothe, 2009)

Les résultats montrent que la combinaison des deux modes (Cconcept) semble efficace ; pour la plupart des requêtes, la combinaison donne la meilleure précision. La précision moyenne en appliquant la combinaison (Ccomb, MAP de 0,193) est supérieure ($> +37\,\%$) à celle obtenue par la recherche en texte libre (Ctexte, MAP de 0,141), et est également supérieure ($> +24\,\%$) à celle obtenue par la recherche par concepts (Cconcept, MAP de 0,155). Suivant les conclusions de Smucker *et al.* (2007), nous avons confirmé ces observations en testant la significativité statistique de la différence entre deux approches pour toutes les requêtes (t-test pairé bilatéral et test de randomisation). D'une part, la différence entre les résultats de Ccomb et de Ctexte est aussi significative ($p < 0,04$) et d'autre part, la différence entre les résultats de Ccomb et de Cconcept est significative ($p < 0,01$). Cependant, pour certaines requêtes un mode unique est plus efficace ; il serait intéressant d'étudier les différents types de requêtes correspondant à chaque mode de recherche à l'instar des études que nous avons menées dans le domaine de la RI XML (cf. chapitre 4, sections 3 et 5).

La seconde étude a été menée pour comparer l'efficacité de notre approche par rapport à une combinaison fondée sur une approche largement utilisée telle que la fonction cosinus :

- Cconcept, Ctexte et Ccomb utilisent notre solution comme décrit précédemment,
- Coscomb résulte de la fusion de résultats en appliquant la méthode CombSUM (Fox et Shaw, 1994) ; les résultats combinés sont obtenus séparément suivant les modes de recherche en texte libre et par concepts en utilisant la fonction cosinus comme fonction d'appariement et une pondération de type *tf.idf* plutôt que notre modèle.

La figure 1.4 compare les processus de recherche suivant notre approche et une approche de type cosinus suivant le critère de rappel/précision.

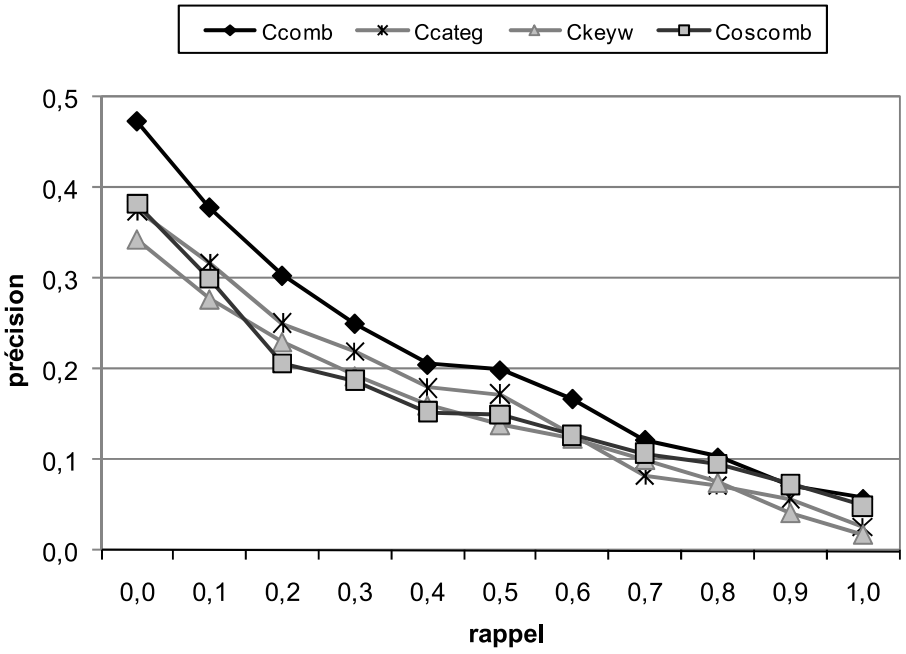


Figure 1.4 – Comparaison d'approches (Hubert et Mothe, 2009)

Les résultats montrent que la combinaison des modes suivant notre approche est plus efficace. Au regard de la MAP, la combinaison utilisant notre approche (Ccomb, MAP de 0,193) conduit nettement à des résultats supérieurs ($> +28\,\%$) à ceux de l'approche cosinus (Coscomb, MAP de 0,150). Cette observation est confirmée par les tests de significativité entre les deux approches pour toutes les requêtes (t-test pairé bilatéral et test de randomisation) indiquant que la différence est significative ($p < 0,02$).

D'autres expérimentations que nous avons menées ont montré que la combinaison des modes fondés sur l'approche cosinus s'avère plus performante que les modes séparés ; ceci renforce l'intérêt de combiner les deux modes de recherche.

3.3 Recherche d'information suivant des ontologies

Le développement récent des ontologies nous a tout naturellement orientés vers leur utilisation dans un processus de recherche d'information. Des travaux pour l'indexation sémantique de documents suivant des ontologies ont été menés (Desmontils et Jacquin, 2002; Pouliquen *et al.*,

2002; Guha *et al.*, 2003; Kiryakov *et al.*, 2004) et notamment au sein de notre équipe (Hernandez *et al.*, 2005). Les domaines comme les corpus évoluent constamment et cette dynamique doit être prise en compte dans un processus de recherche d'information basé sur des ontologies. L'ontologie doit évoluer pour représenter au mieux le domaine correspondant ; cette évolution de l'ontologie a donc une incidence sur les documents déjà indexés. Inversement, l'indexation de nouveaux documents peut amener à faire évoluer l'ontologie ; cette évolution du corpus a donc également une incidence sur l'ontologie définie. La prise en compte de cette dynamique peut se résumer à réappliquer le processus d'indexation complet ou à prendre en compte dynamiquement les évolutions. C'est à ce type de problématique que nous nous intéressons à présent. Des travaux visent ensuite à permettre la recherche d'information à partir d'ontologies en se basant notamment sur des mesures de similarité entre concepts (Andreasen *et al.*, 2003). Cependant, une ontologie n'est pas seulement constituée de concepts, elle comprend également des relations entre ces concepts. La recherche d'information suivant une ontologie n'est plus seulement basée sur des concepts isolés mais plutôt suivant des graphes constitués de concepts et de relations entre ces concepts. C'est à ce type de problématique que nous nous intéressons également au travers de la définition d'une mesure de similarité prenant en compte l'indexation des documents sous forme de tels graphes.

3.3.1 Des hiérarchies de concepts aux ontologies

Comme introduit précédemment, les hiérarchies de concepts, les thésaurus et les ontologies sont des moyens de décrire un domaine. Une hiérarchie de concepts est fondée sur la relation de subsomption « est un » entre concepts. Un thésaurus est un ensemble de termes normalisés et contrôlés, reliés entre eux par des relations sémantiques (hiérarchiques, associatives, ou d'équivalence). Une ontologie définit le sens des objets à travers les mots ou expressions qui les désignent et à travers une représentation structurée ou formelle de leur rôle dans le domaine (Aussenac-Gilles et Mothe, 2004). Ces trois moyens de décrire un domaine peuvent être classés en termes de degré de formalisation et du potentiel d'inférence croissants, des hiérarchies de concepts jusqu'aux ontologies en passant par les thésaurus (Brewster et Wilks, 2004).

Bachimont (2000) précise que le niveau de spécification formelle permettant de restreindre l'interprétation de chaque concept et ainsi d'en donner la sémantique (le « degré d'engagement sémantique ») distingue également les ontologies. Nous nous intéressons aux ontologies terminologiques (van Heijst *et al.*, 1997) qui spécifient les termes utilisés pour représenter la connaissance d'un domaine. Le W3C recommande Web Ontology Language (OWL) (McGuinness et van Harmelen, 2004) comme standard pour représenter les ontologies. OWL se veut plus représentatif du contenu du Web que XML, RDF et RDF-Schéma en apportant un nouveau vocabulaire avec une sémantique formelle.

Des travaux menés au sein de notre groupe se sont intéressés à la transformation de thésaurus normalisés en ontologies. Différentes normes régissent la conception de thésaurus. Elles édictent les règles relatives à l'identification des descripteurs, des relations entre eux et des termes composés, à la présentation et la gestion du thésaurus mais également les modalités de maintenance. Les principales normes internationales sont ISO 2788 :1986 et ANSI/NISO Z39.19-2003 pour les thésaurus monolingues, ISO 5964 :1985 pour les thésaurus multilingues. Pour la France, il existe les normes AFNOR NF Z47-100 à AFNOR NF Z47-103. Nos travaux ont abouti à la définition d'un ensemble de règles pour la transformation de thésaurus normalisés en ontologies ont été définies

dans (Chrisment *et al.*, 2006a,b). Nous avons proposé une mise en œuvre de ces règles pour transformer le contenu d'un thésaurus normé en une ontologie légère de domaine écrite en langage OWL (Hubert *et al.*, 2009b).

3.3.2 Un modèle d'indexation sémantique dynamique

Tout SRI est confronté à un problème de volume d'information à traiter pour répondre à un besoin d'information exprimé par l'utilisateur. Pour pallier cette difficulté des structures de données adaptées telles que les fichiers inverses ont été proposées pour optimiser le processus de recherche d'information. Il s'agit de répondre à des critères de performance comme la qualité des résultats, de temps de réponse aux requêtes, de disponibilité du système. Outre l'indexation initiale d'une collection, l'évolution du corpus implique une ré-indexation ; des structures de données doivent donc être conçues de façon à pouvoir répondre à ce besoin d'évolution. Dans le contexte d'une indexation sémantique suivant des ontologies, la ré-indexation intervient lorsque la collection de documents évolue mais également lorsque l'ontologie utilisée pour l'indexation évolue.

Ainsi, tout l'enjeu est de définir une structure de données qui supporte le processus de recherche tout en permettant les mises à jour de la structure suite aux évolutions de la collection de documents et de l'ontologie.

Un premier objectif est donc, qu'en cas d'évolution des documents, il ne soit nécessaire de ne réindexer que les documents concernés voire de ne modifier l'index que par rapport aux parties de documents modifiées. Le principe est donc de conserver les données nécessaires pour le processus de recherche c'est-à-dire relatives aux éléments de l'ontologie qui décrivent les documents et d'autre part des données permettant une exploitation dynamique de l'index en fonction des évolutions. Ces travaux ont été menés dans le cadre de la thèse de Bachelin Ralalason (2010) co-encadrée avec le professeur Josiane Mothe.

3.3.2.1 Description du modèle d'indexation sémantique dynamique

Le modèle de données présenté dans la figure 1.5 prend notamment en compte le double objectif d'utilisation d'ontologies lors de l'indexation et de la recherche, et l'actualisation dynamique (Hubert *et al.*, 2009a). L'actualisation dynamique réside dans l'exploitation qui est faite du modèle, présentée dans la section 3.3.2.2.

Le modèle reprend l'indexation présentée précédemment pour les UI représentées à l'aide de concepts (cf. section 3.1.2.1). Un document est représenté par un ensemble de termes ou expressions avec un poids associé. Il est également représenté par un ensemble de concepts issus d'une ontologie avec un poids associé. Les relations entre concept identifiées dans le document sont également conservées étant donné que deux concepts peuvent être reliés par différentes relations au sein de l'ontologie. Une occurrence d'une expression est repérée par sa position relative par rapport à un délimiteur du document et peut être reliée à un concept de l'ontologie. Les délimiteurs subdivisent les documents en blocs.

Trois états sont possibles pour un document : *normal* (cas des documents intégrés dans le système avec indexation à jour), *mis à jour* (cas des documents dont le contenu a été modifié depuis la précédente indexation) ou bien *effacé* (le document est retiré logiquement du système depuis la précédente indexation).

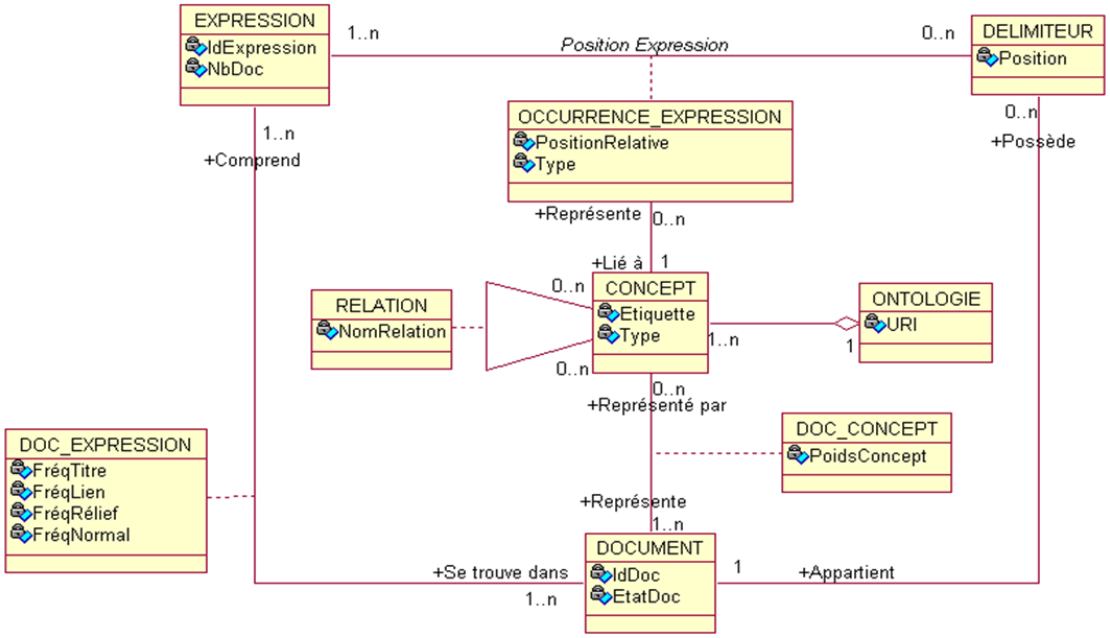


Figure 1.5 – Diagramme de classes représentant les données utilisées pour l'indexation (Hubert *et al.*, 2009a)

3.3.2.2 Exploitation du modèle en réponse à la dynamique de la collection de documents

Le modèle de données sur lequel se base notre approche est axé sur l'utilisation d'ontologies lors de l'indexation et la recherche. De plus, ce modèle permet une actualisation dynamique consécutive à l'évolution de la collection de documents et de l'ontologie. L'actualisation dynamique est liée à une exploitation du modèle orientée dans ce sens (Hubert *et al.*, 2009a). Trois cas sont distingués : l'ajout, la modification et la suppression d'un document.

L'ajout d'un document regroupe : le découpage du document en blocs à l'aide de délimiteurs, l'extraction des expressions, et l'identification éventuelle des concepts et relations entre concepts correspondant aux expressions.

La suppression ou la modification d'un document implique dans un premier temps uniquement la mise à jour de son statut avant de mettre à jour les différentes informations relatives au document dans la base. Pour un document supprimé, son statut est changé en *Effacé* et ce document ne sera plus pris en compte par les requêtes. Après les mises à jour des données du document, le document sera supprimé physiquement de la collection.

À chaque modification d'un document, détectée par exemple à l'aide de la méthode *diff* (Ukkonen, 1985), son statut passe à l'état *Modifié* avant la phase de mise à jour des données relatives aux expressions du document. Les requêtes faisant appel au document en cours de modification utilisent les anciennes valeurs dans la base en attendant la fin des mises à jour des données relatives au document modifié. L'état du document redevient *Normal* après les mises à jour des données. Les modifications apportées à un document peuvent concerner l'ajout, la suppression et la modification d'occurrences d'expressions et sont traitées par bloc. Les traitements sont identiques à ceux appliqués lors de l'ajout et la suppression d'un document. La modification d'une occurrence d'expression est considérée comme une suppression puis un ajout d'expression.

Suite à des ajouts d'expression successifs, la taille du bloc peut devenir trop importante et entrainer l'éclatement du bloc pour réduire le nombre de mises à jour futures. Inversement, la suppression d'expressions peut conduire à des blocs de petite taille et entrainer la fusion de blocs.

3.3.3 Le contexte d'application du projet DynamO

Le projet DynamO (DYNAMic Ontology for information retrieval) doit permettre d'améliorer la recherche d'information et la satisfaction des utilisateurs en prenant en compte la dynamique des documents consultés et celle de la représentation du domaine de référence suivant une ontologie.

Le projet DynamO s'intéresse à l'évolution tant des documents que de l'ontologie représentant le domaine. Ce projet se place dans le contexte où l'afflux de documents, l'insuffisance de l'ontologie pour annoter les documents, une simple évolution de l'ontologie et une réannotation générale ne sont pas concevables. Il s'efforce donc de proposer des solutions pour pallier ces insuffisances pour maintenir une recherche d'information sémantique efficace.

Dans ce contexte et pour prouver le caractère générique des solutions développées au cours du projet DynamO, trois domaines d'application seront expérimentés : l'archéologie des techniques, le diagnostic automobile et la gestion de projets informatiques.

Le projet rassemble principalement deux aspects : la maintenance de l'ontologie et le processus d'annotation et de recherche de documents. Notre groupe s'intéresse à ce deuxième aspect. Notre participation concerne d'une part le pilotage du lot de travail concernant l'élaboration des spécifications du processus d'annotation et de recherche. D'autre part, nous participons à l'élaboration de solutions adaptées en liaison avec notre travaux de recherche et relatives à la définition d'un modèle d'indexation sémantique dynamique, de méthodes d'appariement entre représentations de requêtes et de documents, et de principe de reformulation de requêtes.

4 Bilan et Perspectives

Ce chapitre a détaillé nos contributions concernant la prise en compte du domaine de recherche lors de la RI. Mes travaux se sont intéressés à l'exploitation de représentations de domaine telles que les hiérarchies de concepts, les thésaurus et les ontologies dans le processus de RI.

4.1 Contributions

Une première contribution a été de définir un modèle de RI qui supporte une recherche d'information basée sur une description du domaine sous formes de hiérarchies de concepts. L'objectif de ce modèle a été, d'une part, de permettre la représentation des documents suivant des concepts issus de hiérarchies décrivant un domaine au travers d'un processus de catégorisation. D'autre part, le modèle a été appliqué à l'appariement entre requêtes formulées à partir de concepts et documents décrits par des concepts ; le principe est que les utilisateurs décrivent leurs besoins d'information suivant des concepts des hiérarchies.

De plus, le modèle a été également appliqué à une recherche d'information *ad hoc* en texte libre lorsque l'expression du besoin n'est pas possible à partir de concepts. Outre l'idée d'offrir une possibilité d'interrogation en texte libre ou une possibilité d'interrogation par concepts,

nous avons montré l'intérêt d'exploiter notre modèle pour combiner recherche en texte libre et recherche basée sur une représentation du domaine. Ainsi, les utilisateurs peuvent définir leurs besoins d'information en mêlant concepts identifiés dans le domaine et texte libre, par exemple lorsque l'identification de concepts décrivant la totalité de leur besoin est difficile.

Les propositions ont été mises en œuvre, expérimentées et validées, pour partie, dans le cadre des projets européens IRAIA et eStage. Des expérimentations menées sur des collections de référence c'est-à-dire Reuters-21578 et « TREC9 filtering track » ont permis d'évaluer l'efficacité des différentes propositions.

Dans le domaine de la RI sémantique à base d'ontologies, la seconde contribution a consisté à définir un modèle d'indexation sémantique qui prenne en compte la dynamique de la collection et de l'ontologie. Le modèle supporte l'indexation des documents et des requêtes à l'aide de concepts issus d'une ontologie. Il vise en plus à prendre en compte la dynamique de la collection (ajout, suppression ou modification d'un document) et mettre à jour l'index uniquement sur la partie modifiée.

Le projet ANR DynamO constitue un cadre applicatif à nos propositions et leur poursuite sur les mesures de similarité sémantique.

4.2 Encadrement et diffusion scientifique

Les travaux menés dans cet axe ont donné lieu à différentes publications :

Thème	Publications
Combinaison des recherches par concepts et mots-clés	Revue internationale : – JASIST (Hubert et Mothe, 2009)
Recherche sémantique suivant des ontologies	Revue nationale : – I^3 (Chrisment *et al.*, 2006b) Conférences nationales : – VSST'10 (Dudognon *et al.*, 2010b) – CORIA'09 (Hubert *et al.*, 2009a) – VSST'09 (Hubert *et al.*, 2009b)
Catégorisation automatique de documents	Conférences nationales : – BDA'03 (Augé *et al.*, 2003) – VSST'01 (Augé *et al.*, 2001) Livrables de projet : – IRAIA D3.1.7 (Mothe *et al.*, 2002a) – IRAIA D3.1.8 (Mothe *et al.*, 2002b)
RI géographique et évaluation	Revue nationale : – Document Numérique (Cabanac *et al.*, 2011b) Conférence internationale – ECDL'10 (Palacio *et al.*, 2010b) Conférence nationale – INFORSID'10 (Palacio *et al.*, 2010a)

Une part de ces travaux a permis le déroulement de la thèse de Bachelin Ralalason (2010) que j'ai co-encadrée avec le professeur Josiane Mothe.

Ces travaux ont également conduit à des collaborations avec de nombreux partenaires dans le cadre des deux projets européens IRAIA et eStage ainsi que dans le cadre du projet ANR DynamO.

4.3 Perspectives

Les perspectives que j'envisage à ces travaux sont doubles. Dans le domaine de la recherche sémantique, nous avons fait des propositions pour prendre en compte la dynamique de la collection. Il est nécessaire de compléter ces propositions par une réponse à la dynamique de l'ontologie et à son impact sur l'indexation des documents. La préoccupation est de ne pas remettre en cause totalement l'indexation précédemment réalisée en procédant à une ré-indexation complète. Une orientation à étudier est la gestion de versions. Mes recherches dans le domaine de la gestion de versions pour objets complexes (Hubert, 1997) pourront servir de base à une gestion combinée de versions au niveau de l'index des documents et au niveau de l'ontologie.

Par ailleurs, une seconde perspective concerne la recherche multi-dimensions c'est-à-dire la recherche combinant plusieurs dimensions. Dans ce cadre, nous avons apporté une contribution en proposant un modèle de RI qui combine recherche par concepts et recherche en texte libre. La recherche d'information géographique est un exemple de recherche sur trois dimensions : spatiale, temporelle et thématique. Une première série de travaux initiée récemment a permis de proposer un SRI prenant en compte ces trois dimensions ainsi que de définir un cadre d'évaluation pour les SRI géographiques (Palacio *et al.*, 2010b,a; Cabanac *et al.*, 2011b). Il est maintenant nécessaire d'étudier comment combiner de manière optimale les trois dimensions géographiques. Outre la poursuite de ces travaux dans le domaine de la RI géographique, une perspective à plus long terme est d'étendre les propositions à la prise en compte de dimensions quelconques.

2 Considérer la structure des documents lors de la RI

1 Introduction

La plupart du temps en recherche d'information, les documents sont traités du point de vue de leur contenu textuel. Ce contenu textuel est considéré comme un simple ensemble de mots indépendants.

Au-delà du contenu textuel, des informations spécifiques aux documents existent comme par exemple leur structure en sections et paragraphes. Ces informations ont été longtemps implicites ; une analyse du texte pouvait permettre, par exemple, de retrouver le découpage des documents en phrases en se basant sur la ponctuation voire en paragraphes. Cette information pouvait ainsi ensuite être exploitée.

La structuration implicite est devenue explicite au travers de langages tels que SGML, HTML (Raggett *et al.*, 1998) et XML (eXtensible Markup Language) (Bray *et al.*, 1998) pour décrire les documents. Le nombre de documents construits avec le langage XML ne cesse de croître. De plus en plus d'entreprises migrent leur système de gestion de fonds documentaires vers le langage XML, le standard industriel pour l'échange de données (Fuselier et Chidlovskii, 2006). Considérer la structure des documents en plus du contenu textuel est donc devenu nécessaire pour la RI.

La structuration de ce chapitre est la suivante. La section 2 présente les propositions de la littérature pour traiter les problématiques liées à la RI dans des collections de documents XML et introduit nos propositions. La section 3 détaille nos propositions pour la RI XML. Le bilan de nos contributions et les perspectives de travaux sont présentés dans la section 4.

2 Problématique et travaux de la littérature

Le nombre croissant de collections de documents XML a conduit à la nécessité de développer des systèmes qui permettent la recherche d'information dans de telles collections. Dans ce cadre, les systèmes de recherche XML nécessitent la prise en compte des aspects contenu textuel et structure. En plus de rechercher des documents complets et de laisser l'utilisateur trouver où se trouve l'information qui l'intéresse, la recherche XML offre la possibilité de retrouver uniquement certains éléments au sein des documents ou le meilleur point d'entrée dans un document. La disponibilité de l'information structurelle offre également la possibilité d'exprimer des

indications sur la granularité des éléments retrouvés par le système mais également d'exprimer des indications sur la localisation des concepts recherchés. Des langages d'interrogation tels que NEXI (Trotman et Sigurbjörnsson, 2005) permettent de spécifier de telles indications. Prendre en compte ces aspects constitue une première problématique. L'organisation hiérarchique des éléments dans un document XML et leur imbrication introduit également une autre problématique. En effet, il s'agit de tenir compte du fait que certains éléments sont composés d'autres éléments ayant un impact sur l'estimation de la pertinence des différents éléments par rapport au besoin exprimé par l'utilisateur.

2.1 Évaluation et modèles de RI XML

La création de l'initiative INEX [1] (« Initiative for the Evaluation of XML retrieval ») pour l'évaluation de la recherche d'information XML, soutenue par le réseau d'excellence pour les bibliothèques numériques DELOS [2] a renforcé l'intérêt pour la conception de systèmes de recherche d'information XML. INEX a également introduit de nouvelles problématiques au travers de différentes tâches de recherche telles que la recherche focalisée (« Focused ») qui vise à restituer à l'utilisateur les éléments les « plus appropriés » sans chevauchement des résultats ou la recherche exhaustive (« Thorough ») qui vise à retourner tous les éléments « hautement » pertinents.

De nombreuses propositions de systèmes permettant de rechercher les composants pertinents au sein de documents XML sont apparues principalement au travers des campagnes d'évaluation annuelles INEX. Des études détaillées concernant les propositions en RI XML peuvent être trouvées dans (Tannier, 2006) et (Lalmas, 2009).

Les propositions relatives à la RI XML sont dérivées d'approches définies pour la RI non structurée et étendues pour prendre en compte les aspects introduits par la manipulation de documents XML. Ainsi, il est possible de recenser des propositions en RI XML s'appuyant sur les différents modèles de RI notamment :

- sur un modèle vectoriel (Mass et Mandelbrod, 2005; Crouch *et al.*, 2006; Geva, 2006; Raja *et al.*, 2007),
- sur un modèle de langue (Hiemstra, 2003; Kamps *et al.*, 2004; Ogilvie et Callan, 2006; Ashoori et Lalmas, 2007),
- sur un modèle probabiliste (Larson, 2006; Lu *et al.*, 2006; Gery *et al.*, 2008),
- sur des réseaux bayésiens (Piwowarski et Gallinari, 2005; Zhang et Heng, 2008).

Cependant, les évaluations annuelles INEX de ces propositions ont montré des efficacités très variables de propositions fondées sur le même type de modèle de RI. En effet, les propositions rassemblent des solutions différentes au niveau des composants qui interviennent dans le processus de RI. Ainsi, les extensions des modèles sous-jacents aux différentes approches ont principalement porté sur des propositions au niveau de l'indexation des documents, au niveau de l'appariement requêtes et élément XML et au niveau de la granularité des éléments restitués.

1. http://www.inex.otago.ac.nz
2. http://www.delos.info

2.2 Indexation, appariement et prise en compte de la structure

Au niveau de l'indexation des documents, il est possible de distinguer les approches qui :

- indexent tous les éléments XML (Sigurbjörnsson *et al.*, 2005),
- indexent ou uniquement les éléments feuilles (Geva, 2006),
- excluent certaines balises comme celles relatives à la mise en forme (Kekäläinen *et al.*, 2005),
- indexent seulement certains types d'éléments (balises) issus de la DTD (Gövert *et al.*, 2003; Mass et Mandelbrod, 2005),
- indexent uniquement les éléments au-dessus d'une certaine taille (Hatano *et al.*, 2005).

Au niveau de l'appariement requête et élément XML, les propositions se situent à deux niveaux : les propositions de mesures de similarité qui tiennent compte de la structure des éléments XML, et les propositions d'agrégation de scores suivant la structuration des éléments XML.

Les solutions relatives aux mesures de similarité proposent :

- d'utiliser des valeurs statistiques particulières telles que l'ief (« inverse element frequency ») c'est-à-dire l'inverse du nombre d'éléments XML contenant un terme (Sigurbjörnsson *et al.*, 2005) ou une variante comptant uniquement les éléments de même type (Grabs et Schek, 2002; Theobald *et al.*, 2006),
- d'accorder de l'importance à la taille des éléments XML (Kamps *et al.*, 2004; Ogilvie et Callan, 2006) ou au nombre de changements de sujet (Ashoori et Lalmas, 2007),
- de combiner les valeurs statistiques associées aux termes dans l'élément XML et l'élément racine du document ou de combiner leurs scores (Mass et Mandelbrod, 2006; Sigurbjörnsson et Kamps, 2006). Le même principe est proposé entre éléments et ancêtres autres que racine et parent (Arvola *et al.*, 2005; Lu *et al.*, 2006).

Les propositions autour de l'agrégation de score suggèrent :

- de calculer le score d'un élément en agrégeant les scores de ses enfants en appliquant un facteur de réduction suivant le nombre d'enfants estimés pertinents (Geva, 2006), ou suivant le contexte (Sauvagnat *et al.*, 2006),
- de combiner les valeurs statistiques des termes des éléments et de leur descendants en appliquant un facteur d'augmentation (Gövert *et al.*, 2003),
- pour les approches fondées sur les modèles de langue, de combiner le modèle d'un élément avec celui de son parent et ceux de ses enfants (Ogilvie et Callan, 2003).

La prise en compte des indications de structure a également fait l'objet de propositions. Un type d'approche consiste à comparer les balises de la requête et de l'élément XML (van Zwol, 2006) ou à comparer leurs chemins (Carmel *et al.*, 2003; Popovici *et al.*, 2007). Un autre type d'approche consiste en une interprétation disjonctive de la structure spécifiée dans la requête (Lalmas et Roelleke, 2004). Amer-Yahia *et al.* (2004) applique un principe de relaxation de la requête en construisant un ensemble de sous-requêtes avec des indications de structure moins restrictives puis en fusionnant les résultats obtenus.

Outre la prise en compte des indications de structure exprimée dans la requête par l'utilisateur, une problématique liée à la granularité des éléments du résultat concerne la gestion du chevauchement de certains éléments du résultat dû à l'organisation hiérarchique des documents XML. En effet, renvoyer des éléments qui se chevauchent implique une répétition de contenu.

Par exemple, la tâche de recherche focalisée (« Focused ») proposée dans INEX visant à proposer les éléments pertinents au niveau adéquat de granularité implique une absence de chevauchement de résultat. L'approche la plus répandue consiste à filtrer le résultat afin de conserver, parmi les éléments qui se chevauchent, l'élément ayant le plus haut score de pertinence (Kamps *et al.*, 2007; Popovici *et al.*, 2007; Theobald *et al.*, 2007). L'approche proposée par Mihajlovic *et al.* (2006) consiste à estimer l'utilité d'un élément en fonction de sa pertinence estimée, de sa taille et de ses enfants non pertinents. Dans (Mass et Mandelbrod, 2006), pour choisir au sein d'un document les éléments à conserver parmi ceux qui se chevauchent, des règles sont définies suivant la répartition des descendants pertinents. Clarke (2005) cherche à contrôler le chevauchement en ajustant de manière itérative les scores des éléments contenus ou qui contiennent des éléments mieux classés puis en sélectionnant les n éléments les mieux classés.

2.3 Orientation de nos contributions

Nous nous sommes intéressés à la RI XML aux débuts de l'initiative INEX, c'est-à-dire à la fin de l'année 2002 jusqu'en 2006. Les approches proposées à ce moment-là ont été développées par extension d'un système existant dans le monde du document non structuré. Notre approche a été analogue, notre objectif étant d'étendre l'environnement que nous avions défini précédemment dans le cadre de la recherche d'information dans des collections de documents non structurés (cf. chapitre 1). L'approche retenue a visé l'intégration de chaque aspect lié à la structure (imbrication des éléments, localisation des termes, granularité des éléments du résultat) dans notre modèle de RI. Elle a également visé à ne pas remettre en cause les principes définis pour la recherche sur des documents non-structurés. Nous avons apporté des contributions principalement au niveau de :

- l'indexation des documents et des requêtes,
- l'appariement entre requêtes et éléments XML.

Une autre orientation de nos travaux a consisté à participer à différentes éditions du programme d'évaluation INEX (Pinel-Sauvagnat *et al.*, 2003; Hubert, 2005, 2006, 2007) afin d'évaluer l'efficacité de nos propositions.

3 Contributions pour la RI XML

Nos contributions s'articulent autour de la proposition d'un modèle de RI XML. Cette section présente tout d'abord le principe choisi pour notre approche. Elle détaille ensuite comment les différents aspects introduits par la manipulation de documents XML sont intégrés au modèle de RI proposé. Enfin, elle synthétise les résultats de nos participations à différentes éditions du programme d'évaluation INEX.

3.1 Principe de notre approche de RI XML

Nous avons défini un modèle de recherche au chapitre 1 dans le cadre d'une recherche dans des collections de documents non structurés. Ce modèle a été appliqué à une recherche par catégorie et également une recherche en texte libre.

La manipulation de documents XML introduit de nouveaux aspects à prendre en compte dans

un processus de recherche d'information, liés à l'organisation hiérarchique des documents et à la manipulation de la structure des documents.

Dans le cadre d'une recherche dans des collections de documents non structurés, les documents et les requêtes peuvent être considérés comme des éléments semblables. En effet dans ce cas, ces éléments peuvent être considérés tous les deux comme des ensembles de termes bien que les documents soient en général de plus grande taille que les requêtes. Dans le contexte de la recherche d'information XML, le mélange des aspects liés au contenu textuel et des aspects liés à la structure XML conduit à ce que les requêtes soient très différentes des documents.

Plutôt que de vouloir réduire des éléments différents, c'est-à-dire les besoins d'information et les unités d'information, à un même « moule », comme un sac de mots, notre idée s'attache à prendre en compte leurs spécificités. La comparaison d'éléments semblables ne deviendrait alors qu'un cas particulier parmi d'autres. Une première contribution, présentée au chapitre 1, a été de définir un modèle générique qui puisse être spécialisé et appliqué à chaque cas de figure comme la recherche de documents à partir d'un ensemble de mots-clés et la recherche des catégories représentant un document. Un nouvel objectif est d'appliquer ce modèle générique à la recherche XML à partir d'une requête exprimée en langage NEXI (Trotman et Sigurbjörnsson, 2005).

Le principe est néanmoins de représenter les documents et les requêtes suivant des critères comparables. Ces critères sont d'une part les concepts décrivant les documents et les requêtes, et d'autre part les informations structurelles qui peuvent être associées aux documents ou aux requêtes. Les concepts peuvent être par exemple de simples termes, des expressions composées de plusieurs termes ou des catégories. Les informations structurelles sont par exemple les chemins d'accès à un nœud dans la structure hiérarchique d'un document XML.

La définition générique de la fonction de score présentée au chapitre 1 relative à une recherche sur des documents non structurés est la suivante (cf. chapitre 1, section 3.1.2.2, équation 1.1) :

$$Score(BI_1, UI_1) = \left(\sum_{c \in C_1} imp_{BI}(c, BI_1) \cdot imp_{UI}(c, UI_1) \right) \cdot recouv(BI_1, UI_1)$$

où BI_1 est un besoin d'information et UI_1 est une unité d'information. imp_{BI} dénote une fonction qui estime l'importance du concept c (appartenant à l'ensemble C_1 des concepts de BI_1 et UI_1) dans le besoin d'information BI_1. imp_{UI} dénote une fonction qui estime l'importance du concept c dans l'unité d'information UI_1. *recouv* dénote une fonction qui estime le recouvrement entre BI_1 et UI_1.

Nous avons étendu cette définition pour prendre en compte les aspects structurels à deux niveaux :

- au niveau de la localisation du contenu dans la structure hiérarchique du document,
- au niveau de la granularité d'un élément par rapport à la structure hiérarchique d'un document.

Pour inclure ces deux aspects et par souci de lisibilité deux facteurs liés aux aspects structurels sont ajoutés à la définition générique de calcul de score d'un élément B par rapport à un élément A qui devient la suivante :

$$Score(A,B) = \left(\sum_{c} imp_{BI}(c,A) \cdot imp_{UI}(c,B) \cdot corr_{struc}(c,A,B) \right) \cdot rec_{cont}(A,B) \cdot rec_{struc}(A,B) \qquad (2.1)$$

où A et B sont des éléments (éventuellement représentés différemment)
c est un concept représentant A ou B.

$imp_{BI}(c,A)$	est le facteur qui évalue l'importance du concept c_i pour l'élément A.
$imp_{UI}(c,B)$	est le facteur qui évalue l'importance du concept c_i pour l'élément B.
$corr_{struc}(c,A,B)$	est le facteur évalue la correspondance entre l'information de structure associée au concept c_i dans l'élément A et celle associée à cet élément dans l'élément B.
$rec_{cont}(A,B)$	est le facteur qui évalue l'importance du recouvrement de contenu entre l'élément A et l'élément B.
$rec_{struc}(A,B)$	est le facteur qui évalue l'importance du recouvrement de structure entre l'élément A et l'élément B.

Le score est défini comme l'addition de contributions apportées par les concepts décrivant les éléments comparés. Ce principe permet d'attribuer de la pertinence aussi bien à un élément contenant un seul concept qu'à l'élément contenant tous les concepts. Dans le cas où une importance égale est attribuée aux concepts, la somme implique que les éléments contenant plusieurs concepts sont favorisés par rapport à ceux en contenant moins. Cependant, suivant les définitions des fonctions imp_{BI} et imp_{UI}, c'est-à-dire la manière d'évaluer l'apport d'un concept pour chacun des éléments, un élément contenant un concept particulier peut être favorisé par rapport à un élément contenant plusieurs autres concepts.

La fonction $corr_{struc}$ évalue l'importance du concept c_i vis-à-vis des informations de structure qui lui sont associées dans les A et B.

Enfin, le score est défini suivant les facteurs rec_{cont} et rec_{struc} qui évaluent le recouvrement entre les deux éléments comparés. Le facteur rec_{cont} évalue le recouvrement vis-à-vis du contenu textuel et le facteur rec_{struc} évalue le recouvrement vis-à-vis de la structure. Ces facteurs permettent de rééquilibrer l'apport individuel introduit par chaque concept dans le score attribué à un élément par rapport à l'apport introduit par le nombre de points communs.

Ce principe général de calcul de score permet d'appliquer des mesures de similarité symétriques entre éléments de même nature. Il permet également d'appliquer des mesures spécifiques tenant compte des particularités des éléments comparés et de leur rôle dans la comparaison. En effet, en général en recherche d'information la comparaison s'effectue dans un seul sens. Par exemple, pour la recherche d'information dite *ad hoc* la comparaison effectuée est en général d'un document par rapport à une requête et non l'inverse. La formule de score alors appliquée peut avoir une définition asymétrique.

3.2 Représentation des documents et des requêtes

L'utilisation d'un SRI et l'application d'une fonction de score repose sur un prétraitement des documents et des requêtes. Ce prétraitement communément appelé indexation consiste à extraire des éléments représentatifs et à laisser de côté ceux influençant négativement le processus de comparaison.

Dans le contexte de la recherche d'information XML, nous nous sommes appuyés, pour le processus d'indexation automatique utilisé dans le cadre des documents non structurés (cf. chapitre 1, section 3.1.2.1). Le traitement des documents est étendu pour intégrer les aspects liés à la structure hiérarchique des documents : choix de l'unité d'indexation, conservation des informations de structure. Le traitement des requêtes est étendu à l'extraction des préférences structurelles qu'il est possible d'indiquer dans un contexte XML.

3.2.1 Traitement des documents

Les documents XML sont définis par un ensemble de balises organisées hiérarchiquement. Un document peut être considéré comme un tout, décomposé suivant les balises en différents éléments agrégeant les contenus textuels qu'ils regroupent, ou encore des éléments textuels associés à une organisation hiérarchique. Les différentes balises peuvent également être vues comme définissant des « niveaux de détail » au sein des documents. Il est par exemple possible d'ignorer certaines balises pour la définition de ces niveaux comme les balises de mise en forme. De plus, lorsque les documents sont homogènes, il est possible de ne travailler qu'à certains niveaux pour l'ensemble des documents. Les documents sont homogènes lorsqu'ils partagent des structurations hiérarchiques semblables (mêmes imbrications de balises et mêmes profondeurs d'imbrication).

Contrairement à (Hayashi *et al.*, 2000; Sigurbjörnsson *et al.*, 2005; Mass et Mandelbrod, 2005) qui s'appuient sur plusieurs index correspondant chacun à un niveau différent dans les documents, nous avons choisi de construire un seul index pour les documents au niveau de tous leurs composants possédant un contenu textuel uniquement. Ce choix a été notamment guidé par l'orientation des travaux vers la définition de la fonction de comparaison. En effet, l'indexation au niveau le plus fin représenté par les éléments ayant un contenu textuel est celle qui offre a priori le plus de finesse et donc le plus de possibilités de traitements ultérieurs.

Chaque document est donc parcouru afin d'identifier les balises délimitant un contenu textuel qui définissent les éléments XML que nous manipulons. L'élément XML est identifié par une combinaison de l'identifiant du document dont il est composant et du chemin de localisation de l'élément au sein du document. L'identification de l'élément XML utilise une notation de type Xpath (Clark et DeRose, 1999) étendue pour inclure l'identifiant du document.

Le contenu textuel de chaque élément XML est examiné afin d'en extraire les termes représentatifs et leur nombre d'occurrences dans l'élément. L'extraction de termes met en œuvre notamment la suppression des mots vides en utilisant un anti-dictionnaire. Des traitements supplémentaires comme une racinisation ou une lemmatisation peuvent être également appliqués de manière optionnelle.

Un élément XML peut donc être représenté comme suit en se basant sur la définition d'une UI (cf. chapitre 1, section 3.1.2.1) :

$$E = \left(id, \{(t_1, tf_1), \ldots, (t_n, tf_n)\}\right)$$

où
- id identifie l'élément XML (identifiant du document + chemin de localisation de l'élément dans le document)
- t_i est un terme décrivant l'élément XML
- tf_i est le nombre d'occurrences du terme t_i dans l'élément E.

3.2.2 Traitement des requêtes

Le traitement des requêtes pour la recherche d'information XML regroupe deux aspects : une partie liée au contenu textuel et une partie liée éventuellement à des indications de structure souhaitées. Le traitement de la partie textuelle s'appuie sur les principes du traitement des documents.

Le traitement des indications sur la structure regroupe deux processus. Il s'agit d'une part d'identifier les indications sur la granularité souhaitée par l'utilisateur pour les éléments à retrouver. Ces indications sont a priori indépendantes de l'aspect textuel d'une requête. D'autre part, il s'agit de recenser les indications sur la localisation souhaitée pour les concepts décrivant l'information recherchée. Ces indications sont quant à elles liées aux concepts de la partie textuelle d'une requête. Les indications de structure, qu'elles soient associées aux concepts ou à la granularité des résultats, sont conservées en utilisant une notation de type Xpath (Clark et DeRose, 1999).

Une requête est alors représentée de la façon suivante, en étendant la définition d'un BI (cf. chapitre 1, section 3.1.2.1) aux caractéristiques liées à la structure :

$$Q = \left(rg, \left\{(t_1, tf_1, tp_1, tl_1), \cdots, (t_n, tf_n, tp_n, tl_n,)\right\}\right)$$

où
- rg est l'indication sur la granularité souhaitée pour les éléments XML recherchés,
- t_i est un terme décrivant la requête Q,
- tf_i est le nombre d'occurrences du terme t_i dans la requête Q,
- tp_i est la préférence associée au terme t_i dans la requête Q,
- tl_i est l'indication sur la localisation souhaitée pour le terme t_i dans les éléments XML recherchés.

3.3 Appariement élément – requête

L'appariement entre représentations de documents et représentations de requêtes constitue une étape dans la recherche d'éléments XML répondant à une requête. Dans le cadre de la recherche XML, nous distinguons deux types d'éléments XML : les éléments atomiques qui possèdent un contenu textuel et les éléments composés d'autres éléments XML. L'appariement entre éléments atomiques et requête repose sur l'application d'une fonction de score tenant compte des caractéristiques des éléments XML et des requêtes notamment les termes représentant éléments atomiques et requêtes, les informations de structure associées aux éléments atomiques et les précisions de structure définissant les requêtes.

3.3.1 Appariement avec des éléments atomiques

L'appariement entre éléments atomiques et requête repose principalement sur le contenu textuel. Cet appariement s'appuie sur notre modèle générique en exploitant les caractéristiques des éléments XML et des requêtes. Les préférences sur les concepts recherchés que l'utilisateur peut préciser dans l'expression de son besoin sont prises en compte. La vérification d'une couverture minimale, c'est-à-dire d'une proportion minimum de concepts de la requête dans l'élément XML, complète la définition de l'appariement.

3.3.1.1 *Évaluation de la pertinence du contenu textuel des éléments atomiques*

La correspondance d'un élément XML possédant un contenu textuel par rapport à une requête est évaluée par une fonction calculant un score. La fonction de score pour les éléments atomiques est définie principalement autour des concepts qui apparaissent à la fois dans l'élément XML et la requête.

Compte tenu de la forme générale sur laquelle s'appuie notre SRI (cf. section 3.1), il s'agit donc dans un premier temps d'exploiter les caractéristiques des éléments XML et de la requête pour définir les trois fonctions liées aux concepts partagés, c'est-à-dire :

- la fonction $imp_{BI}()$ qui estime l'importance d'un concept pour la requête,
- la fonction $imp_{UI}()$ qui estime l'importance d'un concept pour l'élément XML,
- la fonction $rec_{cont}()$ qui estime le recouvrement en termes de contenus entre l'élément XML et la requête.

La définition de ces fonctions repose sur les valeurs statistiques communément calculées en recherche d'information comme par exemple le nombre d'occurrences d'un terme ou sa fréquence au sein d'un élément. Le principe de calcul de score que nous avons défini permet de décliner différentes définitions des différentes fonctions constituant le calcul de score dans le but notamment de pouvoir s'adapter à différents corpus.

Les expérimentations réalisées lors de participations à plusieurs campagnes d'évaluation (Pinel-Sauvagnat *et al.*, 2003; Hubert, 2005, 2006, 2007) ont conduit aux définitions suivantes :

- l'importance d'un concept pour un élément XML est donnée par son nombre d'occurrences au sein de l'élément XML. L'hypothèse faite est que plus un concept apparait dans l'élément plus ce concept est important dans la représentation de l'élément. Une autre hypothèse est que l'importance d'un concept pour le besoin d'information dépend de sa sélectivité entre éléments XML constituant le corpus. L'idée reprend celle de l'*idf* en recherche d'information non structurée. Un concept présent dans de nombreux éléments n'est pas discriminant. En revanche, un concept qui apparait beaucoup dans peu d'éléments permet de distinguer ces éléments par rapport aux autres. Les concepts discriminants sont donc plus importants que les autres concepts. Ce principe conduit à définir l'importance d'un concept pour une un élément XML suivant une division par le nombre d'éléments XML du corpus contenant ce concept,
- l'importance d'un concept pour une requête est donnée par sa fréquence au sein de la requête. L'hypothèse faite est que l'importance d'un concept pour un besoin d'information est proportionnelle à son apparition dans la représentation du besoin.,
- enfin, le recouvrement entre une requête et un élément XML est estimé suivant la proportion de termes communs par rapport au nombre maximum de termes communs possibles. L'hypothèse est bien sûr que plus il y a de termes en commun entre deux éléments, plus le recouvrement entre les deux éléments est important.

La base de la définition de la fonction de score entre une requête et un élément XML est la suivante :

$$Score(Q,E) = \left(\sum_i f_{i,Q} \cdot \frac{tf_{i,E}}{ef_i} \right) \cdot \varphi^{\frac{N_{Q,E}}{\min(N_Q, N_E)}} \tag{2.2}$$

Où Q est une requête
E est un élément XML
$f_{i,Q}$ Fréquence du terme t_i dans la requête Q
ef_i Nombre d'éléments XML dans le corpus qui contiennent le terme t_i
$tf_{i,E}$ Nombre d'occurrences du terme t_i dans l'élément XML E
$N_{Q,E}$ Nombre de termes communs à la requête Q et à l'élément XML E
N_Q Nombre de termes distincts dans la requête Q
N_E Nombre de termes distincts dans l'élément XML E
φ Réel positif

Le recouvrement entre requête et élément XML est défini comme une puissance d'un réel positif φ. Ce principe permet de favoriser les éléments XML comprenant de nombreux concepts de la requête par rapport aux éléments comprenant peu de ces concepts. En agissant sur la valeur attribuée à la constante φ il est possible de faire varier l'influence du facteur recouvrement sur le score final par rapport à l'influence des facteurs évaluant l'importance individuelle des concepts.

3.3.1.2 Prise en compte des préférences sur les concepts de la requête

Notre SRI prend en compte la possibilité d'indiquer des préférences sur les concepts recherchés. L'indication de préférences est une notion courante à l'image de la proposition XQUERY Full-Text (Buxton et Rys, 2003) qui préconise les niveaux MUST, SHOULD, MAY pour préciser avec quelle force considérer un concept décrivant la requête. Cette idée est également présente dans des moteurs de recherche tels que Google ou dans la définition de requête de campagnes d'évaluation telle que INEX via l'utilisation de préfixes '+' et '−' pour indiquer (respectivement) les concepts à favoriser et ceux à ignorer. C'est ce type de préférence, considéré comme un souhait et non comme une exigence (comme le serait l'opérateur MUST) que nous avons traité.

La prise en compte des préférences dans notre SRI intervient dans la définition de la fonction de score au niveau de l'évaluation de l'importance d'un concept pour la requête. Il s'agit d'ajouter un poids en fonction du terme de la requête et du type de préférence qui lui est éventuellement associé dans la requête. L'absence d'indication de préférence est considérée comme la préférence par défaut.

Pour intégrer la prise en compte de préférences associées aux concepts décrivant la requête, la fonction de score est étendue pour aboutir à la définition suivante :

$$Score(Q,E) = \left(\sum_i pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E} \right) \cdot \varphi^{\frac{N_{Q,E}}{\min(N_Q, N_E)}} \tag{2.3}$$

Où $pf_{i,Q}$ est une fonction de pondération qui pour un terme donné t_i et une requête Q, renvoie une valeur réelle correspondant au type de préférence associée au concept t_i dans la requête Q. Un des intérêts de cette définition est de pouvoir faire varier la fonction de pondération $pf_{i,Q}$ à chaque application de la fonction de score et notamment pour chaque requête.

3.3.1.3 Vérification d'une couverture minimale de la requête par les éléments atomiques

La notion de couverture introduite dans le modèle de recherche initial détaillé au chapitre 1 reste applicable aux éléments atomiques. Elle complète la fonction de calcul de score afin de considérer dans le processus de recherche uniquement les éléments dont le contenu textuel renferme une part suffisante de concepts de la requête (cf. chapitre 1, section 3.1.3.2). La couverture est intégrée à la fonction de score de la manière suivante :

$$Score(Q,E) = \left(\sum_i pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E} \right) \cdot c_{Q,E} \cdot \varphi^{\frac{N_{Q,E}}{\min(N_Q,N_E)}} \tag{2.4}$$

tel que Si $\frac{N_{Q,E}}{N_Q} \geqslant CT$ Alors $c_{Q,E} = 1,0$ Sinon $c_{Q,E} = 0.0$

où CT Réel positif représentant le seuil tel que $0,0 \leqslant CT \leqslant 1,0$

$N_{Q,E}$ Nombre de termes communs à la requête Q et à l'élément XML E

N_Q Nombre de termes distincts dans la requête Q

N_E Nombre de termes distincts dans l'élément XML E

La couverture vise à assurer que la recherche exploite seulement des éléments où la requête est suffisamment représentée.

3.3.2 Appariement avec des éléments composés

L'appariement entre éléments composés et requête exploite la structure hiérarchique des documents XML. Le principe de cet appariement repose sur l'idée qu'un élément composé rassemble d'autres éléments, atomiques ou eux-mêmes composés. Un élément est ainsi lié à des éléments textuels plus ou moins éloignés. Le principe est donc de calculer le score d'un élément composé à partir des scores calculés pour ses composants atomiques. Il s'agit en quelque sorte de baser l'appariement entre éléments composés et requête sur un principe d'agrégation des scores des éléments atomiques que contiennent les éléments composés.

3.3.2.1 Agrégation de scores pour prendre en compte l'organisation hiérarchique des documents

L'agrégation de scores est une solution appropriée à l'organisation hiérarchique d'un document. Tous les nœuds autres que les feuilles dans l'arborescence d'un document XML peuvent être considérés comme des agrégations d'éléments de niveau inférieur. Une autre approche consisterait à propager l'index des éléments possédant un contenu textuel à travers l'organisation hiérarchique d'un document XML (Cui *et al.*, 2003a). L'index d'un élément composé serait alors construit dynamiquement comme des index des éléments atomiques que contient l'élément composé. La pertinence de l'élément composé peut alors être évaluée comme pour un élément atomique c'est-à-dire possédant un contenu textuel. Néanmoins, considérant la différence entre les éléments composés et les éléments atomiques du point de vue du contenu textuel, nous avons étendu l'évaluation de la pertinence des éléments composés c'est-à-dire en appliquant un principe d'agrégation de scores.

Le principe est de considérer un élément comme pertinent s'il possède un contenu textuel estimé pertinent mais aussi s'il agrège des composants estimés pertinents. Éventuellement, le composé peut-être considéré moins pertinent que son composant. De plus, l'hypothèse considère que plus un composé agrège des composants pertinents plus il est considéré pertinent. L'agrégation de scores permet de traduire ce principe en promouvant les éléments composés de plusieurs éléments estimés pertinents.

Le score d'un élément composé peut donc être défini à partir des scores calculés pour les éléments possédant un contenu textuel dont il est parent. Pour traduire qu'un élément composé est considéré moins pertinent que son composant un coefficient réducteur peut être appliqué. Le principe récursif de l'agrégation conduit à définir le coefficient réducteur comme fonction de la distance entre l'élément textuel et l'élément composé.

Ce principe d'agrégation conduit au calcul de score pour un élément composé selon la définition suivante :

$$Score_a(Q,E) = Score(Q,E) + \sum_l \left((1 - \alpha \cdot \frac{d(E,E_l)}{d(E_r,E_l)}) \cdot Score(Q,E_l) \right) \tag{2.5}$$

où

α (réel) est le coefficient de réduction appliqué entre le score d'un composant et celui de son composé,
E, E_l et E_r sont des éléments XML tels que E_l est un descendant de E possédant un contenu textuel et E_r est l'élément racine parent de l'élément E,
$d(X,Y)$ est la distance entre l'élément X et son descendant (par exemple dans le chemin /article/bdy/sec/p[2], d(bdy, p[2]) = 2).

Ce principe d'agrégation permet de faire varier la propagation des scores des composants textuels vers les éléments parents. Cette variation est obtenue en modifiant la valeur fixée pour le coefficient α.

Cette définition s'applique à tous les éléments y compris ceux qui ne possèdent pas de composant puisque l'on retrouve alors l'application de la fonction de score terminale :

$$Score_a(Q,E) = Score(Q,E)$$

En réalité, nous n'appliquons pas une agrégation de score mais plutôt une agrégation de contributions au score. En effet, nous prenons en compte dans l'élément composé les termes qui contribuent au score de chaque composant. Cette nuance est importante pour la prise en compte des contraintes sur la localisation des concepts recherchés (cf. section 3.4.1). La définition exacte de la fonction de score pour un élément composé est donc :

$$Score_a(Q,E) = Score(Q,E) + \sum_l \left((1 - \alpha \cdot \frac{d(E,E_l)}{d(E_r,E_l)}) \cdot \left(\sum_i pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E_l} \right) \cdot c_{Q,E_l} \cdot \varphi^{\frac{N_{Q,E_l}}{\min(N_Q,N_{E_l})}} \right) \tag{2.6}$$

3.3.2.2 Vérification de la couverture minimale par les éléments composés

La notion de couverture introduite à la section 3.3.1.3 et appliquée lors du calcul de score pour les éléments atomiques s'applique également aux éléments composés. La couverture pour

les éléments composés peut s'appliquer de manière absolue ou relative.

La vérification absolue de la couverture s'applique à un élément composé en considérant la pertinence de ses composants atomiques en tenant compte de la couverture pour ceux-ci. (cf. section 3.3.1.3). La couverture des éléments atomiques se répercute lors de l'agrégation de score au niveau des éléments composés. Dans ce cas, seuls les éléments textuels contenant « suffisamment » de concepts sont pris en compte dans les calculs de score y compris dans les agrégations de scores. L'hypothèse ainsi faite est qu'un élément atomique estimé insuffisamment pertinent ne contribue pas à la pertinence des éléments composés dont il fait partie.

Inversement, la vérification relative de la couverture est appliquée, pour les éléments composés, après agrégation des scores des descendants sans tenir compte de la couverture pour ceux-ci. Avec la vérification relative de la couverture, tous les éléments atomiques sont pris en compte dans les calculs de score y compris dans les agrégations de scores et ne sont conservés ensuite que les éléments contenant « suffisamment » de concepts. L'hypothèse faite est qu'un élément composé peu regrouper « suffisamment » de concepts de la requête bien que ses composants ne contiennent pas individuellement « suffisamment » de concepts. Il suffit que l'union des concepts présents dans les différents composants contienne suffisamment de concepts.

La vérification relative de la couverture est donc intégrée au calcul de score agrégé de la façon suivante :

$$Score_a(Q,E) = c_{Q,E} \cdot \left(Score(Q,E) + \sum_l \left((1 - \alpha \cdot \frac{d(E,E_l)}{d(E_r,E_l)}) \cdot Score(Q,E_l) \right) \right) \tag{2.7}$$

tel que Si $\frac{N_{Q,E}}{\min(N_Q,N_E)} \geqslant CT$ Alors $c_{Q,E} = 1,0$ Sinon $c_{Q,E} = 0,0$

où CT Réel positif représentant le seuil tel que $0,0 \leqslant CT \leqslant 1,0$

$N_{Q,E}$ Nombre de termes communs à la requête Q et à l'élément XML E

N_Q Nombre de termes distincts dans la requête Q

N_E Nombre de termes distincts dans l'élément XML E

Cette définition garde la possibilité de faire varier la valeur de seuil à chaque application de la fonction de score et notamment pour chaque requête (cf. section 3.3.1.3).

3.4 Indications relatives à la structure dans les requêtes

La recherche d'information dans des collections de documents structurés introduit la possibilité de travailler sur la structure des éléments XML. Il est alors possible d'ajouter à l'expression des requêtes des indications relatives à la structure des éléments manipulés et des éléments recherchés.

Les contraintes relatives à la structure XML sont de deux types : les contraintes sur la localisation des concepts recherchés dans la structure hiérarchique des éléments XML et les contraintes sur la granularité des éléments XML souhaités en résultat. Le traitement de ces deux types de contraintes intervient à deux niveaux. Les contraintes liées aux concepts doivent être prises en compte au niveau de la contribution de chaque terme. Les contraintes liées aux éléments XML

doivent être prises en compte au niveau de l'estimation du recouvrement entre élément et requête.

3.4.1 Indications sur la localisation des concepts recherchés

Une indication sur la localisation d'un concept recherché est une indication de structure associée à ce concept dans l'expression de la requête. Cette indication définit une préférence exprimée par l'utilisateur sur la localisation du concept au sein d'un élément XML. Par exemple, un utilisateur recherchera un document qui traite de « construction aéronautique » mais de préférence au niveau du résumé. Ainsi cet utilisateur sera moins intéressé par un document dans lequel « construction aéronautique » apparaît seulement dans une référence bibliographique.

Ce type d'indication est relatif à la correspondance entre localisation indiquée pour un concept donné dans une requête et à la localisation, dans un document, d'un élément XML où apparaît ce concept. Nous avons ainsi considéré que ce type d'indication influait sur la contribution du concept dans le score d'appariement entre un élément XML et une requête.

Ce type d'indication est introduit dans la définition générale (2.1) de la fonction de score (cf. section 3.1), par le facteur $corr_{struc}$ qui estime la correspondance entre l'information de structure associée au concept c dans l'élément A et celle associée à cet élément dans l'élément B.

Pour le calcul de score d'un élément atomique, les indications de localisation des concepts recherche sont prises en compte comme suit :

$$Score(Q,E) = \left(\sum_i ct_{i,Q,E} \cdot pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E} \right) \cdot c_{Q,E} \cdot \varphi^{\frac{N_{Q,E}}{\min(N_Q, N_E)}} \tag{2.8}$$

où $ct_{i,Q,E}$ est une fonction de pondération qui pour un terme donné t_i dans une requête Q et un élément XML E renvoie une valeur réelle qui évalue la correspondance entre la contrainte de localisation associée au terme t_i et le chemin de localisation de l'élément XML au sein d'un document.

Dans un premier temps, nous avons défini une fonction de pondération qui pénalise les éléments dont le chemin de localisation ne correspond pas à la contrainte exprimée sur le terme. En revanche, la fonction de pondération ne modifie pas le score pour les éléments dont le chemin correspond.

La fonction de pondération est la suivante :

$ct_{i,Q,E} = 1 - \beta(1-a)$

tel que

Si l'élément E ne correspond pas à la contrainte sur le terme t_i dans la requête Q
Alors $a = 0$ Sinon $a = 1$
β est un réel positif tel que $0,0 \leqslant \beta \leqslant 1,0$

Un intérêt de cette définition est de pouvoir faire varier simplement la force de la prise en compte des contraintes sur les concepts dans le calcul de score.

La fonction de pondération peut être définie suivant le calcul de ressemblance entre contextes proposé par (Carmel *et al.*, 2003). Ce calcul va au-delà du principe que nous avons choisi en éva-

luant un degré de ressemblance entre structures. Cependant, nous avons fait l'hypothèse qu'un utilisateur s'intéresse davantage à correspondre aux indications de structure qu'il formule plutôt qu'à s'en rapprocher. De plus, notre approche n'élimine pas les éléments dont la structure ne correspond pas aux indications exprimées dans la requête. L'intérêt d'une telle finesse de comparaison des structures doit être étudié par rapport à la correspondance exacte que nous avons choisie. La substitution de notre fonction de pondération par ce calcul reste néanmoins possible. Sa mise en œuvre au sein de notre méthode pourra faire l'objet d'un futur travail.

Les indications de localisation sur les concepts doivent également être prises en compte dans le calcul de scores des éléments composés. En effet, bien que ne possédant pas directement un contenu textuel, un élément composé peut être considéré comme possédant indirectement le contenu textuel des éléments atomiques qu'il regroupe. Ainsi, si un élément composé possède indirectement un concept de la requête et qui correspond à la localisation indiquée pour ce concept, cet élément composé doit être considéré plus intéressant pour l'utilisateur qu'un élément composé qui ne correspondrait pas à cette indication.

Les contraintes sur les termes sont prises en compte pour les éléments composés en étendant le calcul de score agrégé comme suit :

$$Score_a(Q,E) = c_{Q,E} \cdot Score(Q,E) + c_{Q,E} \cdot \sum_l \left((1-\alpha \cdot \frac{d(E,E_l)}{d(E_r,E_l)}) \cdot \sum_i \left(ct_{i,Q,E} \cdot pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E_l} \right) \cdot \varphi^{\frac{N_{Q,E_l}}{\min(N_Q,N_{E_l})}} \right) \quad (2.9)$$

avec $Score(Q,E) = \left(\sum_j ct_{j,Q,E} \cdot pf_{j,Q} \cdot \frac{f_{j,Q}}{ef_j} \cdot tf_{j,E} \right) \cdot c_{Q,E} \cdot \varphi^{\frac{N_{Q,E}}{\min(N_Q,N_E)}}$,

où $ct_{x,Q,E}$ est une fonction de pondération comme défini plus haut.

Cette définition est cohérente pour les éléments atomiques avec celle donnée plus haut puisque dans ce cas $Score_a(Q,E) = c_{Q,E} \cdot Score_t(Q,E)$

3.4.2 Indications sur la granularité des résultats

Une indication sur la granularité des éléments à retourner comme résultat est une indication de structure complémentaire dans l'expression de la requête. Cette indication exprime une préférence de l'utilisateur vis-à-vis de la localisation des éléments recherchés au sein des documents XML. Par exemple, un utilisateur recherchera de préférence les résumés des documents et sera moins intéressé par les références bibliographiques.

Ce type d'indication de structure n'est pas lié aux concepts exprimés dans la requête, il intervient à un niveau plus global. Il influence globalement la pertinence estimée suivant le besoin exprimé du point de vue des concepts. Nous avons donc choisi de le prendre en compte a posteriori du calcul de score défini précédemment (cf. section 3.4.1) basé sur la partie de la requête liée aux concepts recherchés et aux indications de localisation associées.

Ainsi pour le calcul de score pour les éléments composés, le traitement des indications de granularité des résultats est intégré de la façon suivante :

$$Score_a(Q,E) = cg_{Q,E} \cdot c_{Q,E} \cdot Score(Q,E) + cg_{Q,E} \cdot c_{Q,E} \cdot \sum_l \left((1 - \alpha \cdot \frac{d(E,E_l)}{d(E_r,E_l)}) \cdot \sum_i \left(ct_{i,Q,E} \cdot pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E_l} \right) \cdot \varphi^{\frac{N_{Q,E_l}}{\min(N_Q,N_{E_l})}} \right) \quad (2.10)$$

où $cg_{Q,E}$ est une fonction de pondération qui, pour une requête Q et un élément XML E, renvoie une valeur réelle qui évalue la correspondance entre la contrainte de granularité associée à la requête et le chemin de localisation de l'élément XML au sein d'un document.

À l'image de la fonction de pondération définie pour les contraintes sur les termes, nous avons défini dans un premier temps une fonction de pondération simple. Celle-ci pénalise les éléments dont le chemin de localisation ne correspond pas à la contrainte exprimée. En revanche, la fonction de pondération ne modifie pas le score pour les éléments dont le chemin correspond. La fonction de pondération est la suivante :

$cg_{Q,E} = 1 - \gamma(1-a)$

tel que

Si l'élément E ne correspond pas à la contrainte de granularité de la requête Q

Alors $a = 0$ Sinon $a = 1$

γ est un réel positif tel que $0,0 \leqslant \beta \leqslant 1,0$

Cette définition est également valide pour les éléments atomiques puisque dans ce cas $Score_a(Q,E) = cg_{Q,E} \cdot c_{Q,E} \cdot Score(Q,E)$

Un intérêt de cette définition est de pouvoir faire varier simplement la force de la prise en compte des contraintes de granularité dans le calcul de score.

À l'instar de la prise en compte des indications sur la localisation des concepts recherchés, notre fonction de pondération peut également être substituée par un calcul de ressemblance entre structures comme celui proposé par (Carmel *et al.*, 2003) (cf. section 3.4.1).

3.5 Le contexte des projets Ws-Talk et Quest

Cette section présente les deux projets qui ont constitué des cadres applicatifs de nos propositions.

D'une part, le projet Ws-Talk (Web services communicating in the language of their user community) a eu pour but de de combler le fossé qui existe entre le langage naturel quotidien utilisé pour décrire les besoins d'informations et celui utilisé pour manipuler nombre de services Web comme par exemple le langage XML. L'objectif a été le développement de systèmes permettant de traduire une requête du client en forme compréhensible par une machine, de localiser les services Web appropriés, de les interroger et de renvoyer les résultats à l'utilisateur. Le projet Ws-Talk s'est intéressé à la définition d'une interface assurant une communication interservices via un langage sémantique structuré. Il a proposé de créer un niveau sémantique qui permet d'obtenir des vues sur les services Web, exprimé en langage naturel. Ainsi, l'objectif a été de construire des applications interopérables et interconnectées et de permettre aux fournisseurs et consommateurs de services de définir et implanter de nouveaux services ou d'adapter les services existants.

D'autre part, le projet QUEST (QUEry reformulation for STructured document retrieval) a eu pour objectif de concevoir, implanter et évaluer des mécanismes pour la reformulation de requêtes pour une recherche de documents structurés efficace. Ces mécanismes combinent contenu, structure et connaissance sémantique pour aider les utilisateurs à accéder à des collections volumineuses de documents structurés. Les mécanismes ont été appliqués à des documents XML, langage de structuration de données adopté par le W3C. L'aspect contenu n'est pas spécifique à XML puisque tout document possède un contenu mais la structuration XML peut entrainer une vision composite des contenus. L'aspect structurel est dû au fait que les composants de document XML sont liés via une structure logique. Enfin, l'aspect sémantique provient des attributs c'est-à-dire des métadonnées qui peuvent être associés aux éléments XML. Un mécanisme de reformulation de requête pour XML doit considérer ces trois aspects ainsi que leurs interrelations. Le projet a rassemblé différents objectifs notamment le développement de mécanismes basés sur les principes de reformulation de requête traditionnels qui peuvent être appliqués ou adaptés à XML et d'autres qui réclament un développement spécifique à XML. Un autre objectif a été l'évaluation de l'efficacité des mécanismes développés sur de grandes collections telles que celles proposées par les campagnes d'évaluations INEX.

3.6 Participations aux campagnes INEX

INEX est définie comme une initiative pour l'évaluation des systèmes de recherche XML. Cette initiative a été lancée en 2002 et a subi différentes évolutions depuis son lancement. Cette initiative constituait un cadre approprié pour mener différentes expérimentations afin d'évaluer l'apport de nos propositions. Notre groupe a participé aux campagnes proposées de 2003 à 2006 (incluse). Ces participations ont eu différents objectifs notamment liées aux évolutions apportées aux différentes campagnes.

3.6.1 Cadre INEX de 2002 à 2006

INEX fournit des jeux de tests constitués d'un corpus de documents XML, de requêtes orientées vers la recherche XML et des réponses associées (jugements de pertinences). INEX fournit également des méthodes d'évaluation des résultats fournit par un système.

Les campagnes INEX permettent aux participants d'évaluer leur système suivant les critères définis chaque année. Les campagnes INEX permettent également aux participants de comparer l'efficacité de leur système par rapport aux autres sur une base commune. Bien que n'ayant pas un caractère absolu, ces évaluations permettent néanmoins d'obtenir un éclairage sur les performances des systèmes dans un cadre donné.

Différentes tâches ont été proposées depuis le début de l'initiative. Compte tenu de notre objectif visant à proposer un SRI pour chercher des éléments XML textuels répondant à un besoin exprimé par l'utilisateur, nos participations se sont focalisées sur la tâche principale « *ad hoc* ». L'objectif de cette tâche est évaluer la capacité des systèmes à répondre à des requêtes sur le contenu textuel seul ou des requêtes combinant contenu textuel et contraintes sur la structure.

Le corpus de documents utiliséutilisé jusqu'en 2004 contient environ 12 000 articles de journaux scientifiques publiés par « IEEE Computer Society » entre 1995 et 2002. Tous les documents sont structurés suivant la même DTD. Ce corpus rassemble plus de 8 millions d'éléments XML de

tailles et granularités différentes. Ce corpus a été enrichi en 2005 par les articles publiés jusqu'en 2004. Le corpus de l'édition 2005 d'INEX regroupe plus de 11 millions d'éléments XML.

Jusqu'en 2005 INEX définit deux types de requêtes :

– les requêtes de type CO (« Content Only ») qui précisent le besoin d'information uniquement du point de vue contenu textuel,
– les requêtes de type CAS (« Content And Structure ») qui précisent le besoin d'information du point de vue contenu textuel et du point de vue structure XML.

Les deux types de requêtes sont issues de descriptifs de besoin d'information définis par un titre, une description brève, une description détaillée et des mots-clés. Le titre correspond à la requête soumise au SRI pour la tâche *ad hoc*. Pour les requêtes de type CAS, le titre inclut des indications de structure sur les concepts recherchés et la granularité des résultats. Ces indications de structure sont précisées suivant une notation de type XPath (Clark et DeRose, 1999). Le langage NEXI (Trotman et Sigurbjörnsson, 2005) a ainsi été défini pour exprimer les requêtes de type CAS. L'utilisation des autres parties descriptives des besoins d'information a également été autorisée pour des soumissions complémentaires de résultats.

Les requêtes fournies pour les évaluations durant chaque campagne INEX sont sélectionnées parmi des requêtes candidates proposées par les participants à la campagne.

La pertinence est définie suivant deux dimensions : l'exhaustivité et la spécificité. L'exhaustivité traduit le degré avec lequel un élément couvre la requête. La spécificité traduit le degré avec lequel un élément traite de toute l'information recherchée.

Les jugements de pertinence sont réalisés par les participants auteurs des requêtes. Une sélection issue des résultats soumis par tous les participants est soumise au participant ayant proposé la requête. À l'aide d'une interface graphique appropriée, ce dernier examine les résultats et établit la pertinence des éléments XML suivant les deux dimensions. Ces jugements de pertinence servent ensuite de référence pour les évaluations.

3.6.1.1 Caractéristiques des éditions INEX jusqu'en 2004

Jusqu'à l'édition INEX 2004 incluse, les évaluations ont été basées sur les notions de précision et rappel en tenant compte du degré de pertinence des éléments XML suivant les deux dimensions. Les mesures « officielles » utilisées jusqu'à l'édition 2004 d'INEX (de Vries *et al.*, 2004) sont basées sur la méthode precall (Raghavan *et al.*, 1989) et la probabilité P(rel|retr) qu'un document vu par l'utilisateur soit pertinent.

Des fonctions dites de « quantification » ont été définies pour transformer le degré de pertinence exprimé suivant les deux dimensions (exhaustivité et spécificité) en un degré de pertinence suivant une échelle unique. Différentes fonctions de quantification ont été proposées pour décrire différentes préférences utilisateur. Par exemple, la quantification *strict* est appliquée pour évaluer la capacité des systèmes à retrouver les éléments XML totalement exhaustifs et entièrement spécifiques. La quantification *generalised* est appliquée pour évaluer la capacité des systèmes à retrouver tous les éléments pertinents même partiellement. Un score agrégé a également été défini comme la moyenne des différents scores d'évaluation obtenus avec les différentes quantifications.

Les évaluations ont été séparées en deux groupes :

– l'évaluation des requêtes de type CO,

– l'évaluation des requêtes de types CAS. Dans ce cas, les contraintes de structure ont été considérées comme de vagues conditions. Les éléments du résultat n'étaient pas tenus de vérifier de manière stricte les conditions exprimées dans la requête.

3.6.1.2 Évolutions apportées au cadre INEX lors de l'édition 2005

L'édition INEX 2005 a connu l'enrichissement du corpus par de nouveaux documents, et a été marquée par le changement des mesures d'évaluation.

L'enrichissement du corpus a été réalisé en deux fois : une première version du corpus a intégré une partie des nouveaux documents puis une deuxième version complète a été fournie. Cette succession de versions du corpus a introduit un biais dans les évaluations puisqu'une partie des résultats officiels soumis par certains participants (dont notre groupe) ont été réalisés sur la première version incomplète du corpus tandis que d'autres résultats officiels soumis par les autres participants ont été construits à partir de la seconde version complète du corpus.

Le changement de mesures d'évaluation a été justifié par le fait que les mesures précédentes ne prenaient pas en compte le recouvrement des éléments pertinents dans les jugements de pertinence servant de référence. Cette lacune pourrait en théorie entrainer des distorsions dans les résultats d'évaluation, à condition toutefois de construire les résultats en ce sens. Les mesures « officielles » d'évaluation depuis l'édition INEX 2005 sont les mesures XCG (Kazai et Lalmas, 2006). Les mesures XCG (eXtended Cumulated Gain) étendent les mesures basées sur le gain cumulé (« Cumulated Gain ») (Järvelin et Kekäläinen, 2002). Elles regroupent les mesures nxCG (« normalized extended cumulated gain ») orientées utilisateur et les mesures ep/gr (« effort-precision/gain-recall ») orientées système. Trois fonctions de quantification sont définies notamment les quantifications *strict* et *generalised* reprenant les mêmes significations que pour les éditions précédentes.

Du point de vue des requêtes, une nouveauté a été introduite dans la définition de certaines requêtes de type CO par l'ajout d'une partie correspondant à une reformulation de la requête CO en y incluant des précisions de structure.

Les évaluations ont été séparées en trois groupes :

– l'évaluation des requêtes de type CO,
– l'évaluation nommée CO+S des requêtes de type CO suivant la reformulation incluant des indications de structure,
– l'évaluation des requêtes de types CAS.

L'évaluation CO regroupe trois sous-tâches : *Focused*, *Thorough* et *Fetch&Browse*. La sous-tâche *Focused* s'intéresse à la capacité des systèmes à retrouver les éléments hautement spécifiques. La sous-tâche *Thorough* qui évalue la capacité à retrouver tous les éléments pertinents. Elle reprend le principe des évaluations CO des éditions précédentes. Enfin la sous-tâche *Fetch&Browse* mélange recherche au niveau document et recherche d'éléments dans ces documents.

L'évaluation CO+S reprend les sous-tâches de l'évaluation CO mais en utilisant des résultats de recherches basées sur les formulations incluant des indications de structure.

L'évaluation CAS regroupe quatre sous-tâches : VVCAS, SVCAS, VSCAS et SSCAS. Ces sous-tâches correspondent aux différentes possibilités de vérifier les contraintes de granularités sur les éléments résultats c'est-à-dire de manière stricte (SxCAS) ou vague (VxCAS) et de vérifier les

contraintes sur le contenu textuel de manière stricte (ySCAS) ou de manière vague (yCAS). La sous-tâche VVCAS qui considère les deux types de contraintes de manière vague reprend donc la philosophie de l'évaluation des requêtes de type CAS des éditions précédentes.

3.6.1.3 Évolutions apportées au cadre INEX lors de l'édition 2006

La campagne INEX 2006 a été marquée par le changement de corpus et l'introduction de mesures d'évaluation supplémentaires liées à l'introduction de nouvelles sous-tâches. Le corpus proposé depuis l'édition 2006 est la collection de documents en anglais issue de Wikipedia. Le corpus rassemble un peu moins de 660 000 articles soit plus de 100 millions d'éléments XML couvrant plus de 100000 catégories de la taxonomie de Wikipedia.

Un autre changement important est la disparition de l'évaluation de type CAS. En revanche, l'ajout d'une formulation incluant des indications de structure a été généralisé à toutes les requêtes de type CO. L'évaluation CO+S des requêtes CO suivant la formulation avec contraintes a fusionné avec l'évaluation CO des requêtes uniquement sur le contenu textuel.

Deux sous-tâches ont été introduites : *AllInContext* et *BestInContext*. La sous-tâche *AllInContext* reprend la philosophie de la tâche *Fetch&Browse* de l'édition précédente en combinant recherche de documents puis recherche d'éléments dans ces documents. La sous-tâche *BestInContext* consiste à chercher l'élément le plus adapté à restituer à l'utilisateur parmi des éléments qui se recouvrent. Des mesures différentes ont été utilisées pour les évaluations des quatre sous-tâches. La mesure ep/gr a été utilisée pour l'évaluation de la sous-tâche *Focused* et la mesure *nxCG* pour la sous-tâche *Thorough*. La mesure *gP* (« generalised Precision ») basée sur les notions de précision et rappel a été utilisée pour l'évaluation de la sous-tâche *AllInContext*. Les mesures *BEP-D* (« Best Entry Point Distance ») ensembliste et *EPRUM* (« expected precision-recall with user modelling ») extension des calculs de précision et rappel ont été appliquées pour la sous-tâche *BestInContext*.

3.6.2 Résultats de nos participations aux campagnes INEX

Les résultats présentés sont extraits de nos participations à INEX de 2003 à 2006. Notre approche a peu évolué durant ces participations. La participation à INEX 2004 a conduit à une configuration de base de notre SRI utilisée dans les éditions suivantes. Un objectif a été ensuite de mesurer l'impact des changements apportés au cadre INEX sur notre méthode en termes de collections de documents, de mesures d'évaluation, de tâches.

3.6.2.1 Résultats liés à notre participation à l'édition INEX 2003

L'édition INEX 2003 a été l'occasion de mettre en place l'adaptation de notre SRI au cadre INEX. L'adaptation s'est concentrée sur le processus d'indexation des documents et le calcul de score pour les éléments atomiques, c'est-à-dire possédant un contenu textuel. La prise en compte de l'organisation hiérarchique des documents c'est-à-dire au travers du calcul de score pour les éléments composés n'a pu être intégrée pour cette édition INEX 2003. Ce manque explique les faibles résultats lors des évaluations officielles.

3.6.2.2 Résultats liés à notre participation à l'édition INEX 2004

La participation à la campagne INEX 2004 a eu pour principal objectif d'étudier le comportement de notre SRI dans sa première version face au cadre défini par INEX et à ses critères d'évaluation. En effet, notre méthode ayant été définie indépendamment des critères retenus dans INEX, son comportement dans le cadre INEX n'était pas réellement prévisible. L'objectif a donc été de mesurer l'efficacité de notre approche face aux critères définis pour la RI XML.

Le tableau 2.1 résume les évaluations des résultats soumis lors de la campagne officielle INEX 2004 pour les requêtes de type CO et les requêtes de type CAS ainsi que des résultats construits a posteriori (étiquetés « postRun ») sur les données de la campagne INEX 2004. Les résultats a posteriori ont été construits en faisant varier les paramètres du SRI tels que : le coefficient de couverture, le coefficient d'agrégation de score, ou le coefficient de recouvrement (cf. section 3.3). La mesure indiquée est la mesure agrégée c'est-à-dire la moyenne des scores d'évaluations suivant les différentes configurations.

Soumission («Run»)	Mesure agrégée	Rang
RunCO2004	0,0783	13/70
postRun1CO2004	0,1069	
postRun2CO2004	0,1135	
Meilleur système CO	*0,1437*	*1/70*
RunCAS2004	0,0784	5/51
postRunCAS2004	0,0970	
Meilleur système CAS	*0,1260*	*1/51*

Tableau 2.1 – Évaluations officielles INEX 2004 suivant la mesure agrégée

Ces premiers résultats ont montré un classement dans le premier cinquième pour la tâche CO et le premier dixième pour la tâche CAS par rapport aux autres systèmes. Le traitement des contraintes a semblé intégré de manière appropriée dans notre SRI au regard de l'évaluation pour les requêtes CAS notamment par rapport à la meilleure évaluation obtenue par les participants. Les résultats construits a posteriori montrent que l'ajustement de paramètres du SRI permet d'améliorer son efficacité suivant la mesure agrégée. L'étude menée a posteriori a donc abouti à l'identification de la configuration donnant la meilleure évaluation au regard des données de la campagne INEX 2004 et que nous avons réutilisée comme configuration de base pour la campagne de l'année suivante.

Au-delà d'une évaluation globale de notre approche, INEX 2004 nous a fourni des évaluations plus détaillées par rapport à différentes quantifications. Ces évaluations nous ont renseignés notamment sur la capacité de notre SRI à retrouver tous les éléments pertinents (quantification *generalised*) et à retrouver les éléments fortement exhaustifs et spécifiques (quantification *strict*).

Le tableau 2.2 synthétise les évaluations détaillées pour les requêtes CO.

	strict		generalised	
Soumission («Run»)	Précision moyenne	Rang	Précision moyenne	Rang
RunCO2004	0,0778	18/70	0,0683	14/70
postRun1CO2004	0,1432		0,0569	
postRun2CO2004	0,1307		0,0903	
Meilleur système CO	*0,1524*	*1/70*	*0,1308*	*1/70*

Tableau 2.2 – Évaluations INEX 2004 suivant les quantifications strict et generalised pour les requêtes CO

Ces résultats montrent également un classement de notre système dans le premier cinquième malgré des résultats en deçà des meilleures évaluations. Les résultats, notamment ceux construits a posteriori, montrent une différence des évaluations entre les deux quantifications. Ceci suggère que la prise en compte des contraintes est plus adaptée à la sélection des éléments fortement exhaustifs et spécifiques. Les évaluations suivant les autres quantifications ont cependant montré une tendance de notre méthode à privilégier les éléments exhaustifs plutôt que spécifiques. Ceci s'explique par la définition de notre moteur suivant une somme de contributions des termes de la requête. Plus l'élément possède des concepts de la requête, c'est-à-dire plus il est exhaustif, plus le score attribué est élevé.

De même, le tableau 2.3 synthétise les évaluations détaillées pour les requêtes CAS.

	strict		generalised	
Soumission («Run»)	Précision moyenne	Rang	Précision moyenne	Rang
RunCAS2004	0,1053	5/51	0,0720	6/51
postRunCAS2004	0,1173		0,1002	
Meilleur système CAS	*0,1375*	*1/51*	*0,1167*	*1/51*

Tableau 2.3 – Évaluations INEX 2004 suivant les quantifications strict et generalised pour les requêtes CAS

Ces résultats montrent également des résultats plus proches des meilleures évaluations pour les requêtes CAS. Ils suggèrent donc une bonne intégration de la gestion des contraintes dans notre SRI. L'évaluation suivant la quantification *strict* reste également supérieure à celle suivant la quantification *generalised* comme précédemment pour les requêtes CO.

Les résultats pour les requêtes CO (Tableau 2.2) ainsi que pour les requêtes CAS (Tableau 2.3) ont également introduit l'idée qu'une seule et même configuration n'est pas adaptée à toutes les quantifications.

3.6.2.3 Résultats liés à notre participation à l'édition INEX 2005

Pour la campagne INEX 2005 nous avons pris comme base la configuration de notre SRI ayant fourni les meilleures évaluations en moyenne (mesure agrégée) sur les jeux de test de 2004. Cette configuration a été définie en faisant varier les paramètres du SRI tels que : le coefficient de couver-

ture, le coefficient d'agrégation de score, ou le coefficient de recouvrement (cf. section 3.3) comme indiqué précédemment. Cette configuration de base a été appliquée pour construire les résultats soumis pour les trois évaluations de type CO, CO+S et CAS.

Le tableau 2.4 présente un extrait des résultats officiels obtenus pour la campagne INEX 2005.

	strict		generalised	
Soumission («Run»)	ep/gr (MAP)	Rang	ep/gr (MAP)	Rang
RunCO2005v1.Focused	0,0332	10/44	0,0477	24/44
RunCO2005v2.Focused	0,0399	9/44	0,0616	20/44
postRun1CO2005.Focused	0,0447			
postRun2CO2005.Focused			0,0842	
Meilleur système CO.Focused	*0,0741*		*0,0917*	
RunCO2005v1.Thorough	0,0129	31/55	0,0490	29/55
RunCO2005v2.Thorough	0,0168	28/55	0,0566	24/55
postRun3CO2005.Thorough	0,0346			
postRun4CO2005.Thorough			0,0708	
Meilleur système CO.Thorough	*0,0497*		*0,0867*	
RunCO2005v1.FetchBrowse	0,0139	12/40	0,0751	14/40
Meilleur système CO.FetchBrowse	*0,0167*		*0,1118*	

Tableau 2.4 – Évaluations INEX 2005 suivant les quantifications strict et generalised pour les requêtes CO

Les évaluations des soumissions officielles montrent des performances variables d'une sous-tâche à une autre et d'une quantification à l'autre. Néanmoins les évaluations de notre SRI restent dans la première moitié voire le premier quart par rapport aux autres participants notamment pour la quantification *strict* excepté pour la sous-tâche *Thorough*. Les soumissions construites à partir de la version complète de la collection (étiquetés *v2*) donnent de meilleurs résultats, la première version n'aurait donc pas dû être proposée.

De nouvelles expérimentations réalisées ensuite avec des configurations différentes de notre SRI (étiquetées *postRun*), en faisant varier principalement le coefficient d'agrégation de score, ou le coefficient de recouvrement (cf. section 3.3), conduisent à de meilleurs résultats. Ces configurations plus efficaces sont différentes d'une sous-tâche à l'autre. Ce constat nous a donc suggéré que différentes configurations de notre SRI étaient adaptées aux différentes sous-tâches et quantifications définies dans INEX. C'est dans ce sens que s'est orienté notre travail y compris pour la participation à INEX 2006.

Les résultats pour les requêtes de type CAS sont résumés dans le tableau 2.5. Les différentes sous-tâches définies pour ce type de requêtes ont permis de mesurer l'efficacité des traitements liés aux contraintes de structure et leur intégration dans notre SRI.

	strict		generalised	
Soumission («Run»)	ep/gr (MAP)	Rang	ep/gr (MAP)	Rang
Run2005v1.VVCAS	0,0379	10/28	0,0804	7/28
Run2005v2.VVCAS	0,0597	3/28	0,0962	5/28
Meilleur système VVCAS	*0,0614*		*0,1283*	
Run2005v1.VSCAS	0,0287	12/23	0,0908	5/23
Run2005v2.VSCAS	0,0386	7/23	0,1057	5/23
Meilleur système VSCAS	*0,0765*		*0,1478*	
Run2005v1.SVCAS	0,0938	4/23	0,1233	3/23
Run2005v2.SVCAS	0,1087	4/23	0,1365	3/23
Meilleur système SVCAS	*0,1762*		*0,1860*	
Run2005v1.SSCAS	0,0504	18/25	0,1219	13/25
Run2005v2.SSCAS	0,0648	14/25	0,1289	12/28
Meilleur système SSCAS	*0,1334*		*0,2297*	

Tableau 2.5 – Évaluations INEX 2005 suivant les quantifications strict et generalised pour les requêtes CAS

Les résultats montrent qu'à l'exception de la tâche SSCAS notre système se situe parmi les systèmes obtenant les meilleures évaluations. Les résultats pour la sous-tâche VVCAS confirment ceux obtenus lors de l'édition précédente pour la tâche analogue (tâche VCAS de INEX 2004). Une explication de résultats plus faibles pour la tâche SSCAS peut être que le SRI utilisé ne permettait pas de prendre en compte complètement les contraintes de structure sur le contenu. La vérification stricte de ces contraintes n'était pas possible. Ce constat a conduit à faire évoluer la prise en compte de ce type de contraintes. La prise en compte de ce type de contraintes telle que définie à la section 3.4.1 a été intégrée pour la participation à l'édition 2006 d'INEX.

De plus, comme pour les requêtes de type CO les soumissions réalisées sur la version incomplète de la collection (étiquetés *v1*) fournissent des résultats inférieures à celles construites sur la version complète de la collection (étiquetés *v2*).

Au regard des expérimentations, l'introduction de sous-tâche a eu plus d'impact que le changement de mesures d'évaluation même si les résultats ont été dégradés pour la tâche *Thorough* déjà présente en 2004. Notre travail notamment pour la participation à INEX 2006, s'est alors orienté vers la construction de résultats en appliquant des configurations différentes de notre SRI suivant les sous-tâches d'évaluations.

3.6.2.4 Résultats liés à notre participation à l'édition INEX 2006

La participation à l'édition INEX 2006 avait deux objectifs :

– mesurer l'impact du changement de collection sur les performances de notre SRI. La collection IEEE s'inscrivait dans le domaine de l'informatique. La nouvelle collection issue de Wikipedia est plus « généraliste » et donc multi-domaines.

– évaluer l'impact du paramétrage de notre SRI en fonction de la tâche et de la quantification.

Les résultats présentés sont ceux pour les tâches *Thorough* (Tableau 2.6) et *Focused* (Tableau 2.7) déjà présentes dans les éditions précédentes afin de permettre la comparaison. Les résultats pour la nouvelle tâche *BestInContext* sont également présentés (Tableau 2.8) pour mesurer l'adaptation de notre SRI à une nouvelle tâche plus spécifique.

Thorough	generalised	
Soumission («Run»)	ep/gr (MAP)	Rang
RunCO2006.Thorough	0,0147	47/106
RunCAS2006.Thorough	0,0161	40/106
postRunCO2006.Thorough	0,0259	
Meilleure évaluation	*0,0384*	

Tableau 2.6 – Évaluations INEX 2006 pour la tâche Thorough suivant la quantification generalised

Les résultats officiels restent dans la même proportion par rapport à l'ensemble des soumissions (notre rang a augmenté mais il est à rapprocher du nombre de soumissions qui a quasiment doublé par rapport à l'année précédente). Le résultat construit a posteriori, basé sur une configuration différente montre une évaluation nettement supérieure (> + 60 %). Il paraît donc important d'étudier l'identification de configurations du SRI adaptées aux sous-tâches. Par ailleurs, les résultats indiquent que le changement de collection ne semble pas affecter les performances de notre SRI.

Par ailleurs, l'utilisation de la formulation de requête avec contraintes de structure aboutit à de meilleurs résultats. Ceci confirme le traitement approprié des contraintes de structure par notre SRI. L'intérêt d'inclure des indications sur la structure est également souligné par ces résultats.

Focused	generalised					
Soumission («Run»)	nxCG@10	Rang	nxCG@25	Rang	nxCG@50	Rang
Run1CO2006.Focused	0,2472	42/85	0,1905	43/85	0,1558	36/85
Run2CO2006.Focused	0,2435	46/85	0,1843	48/85	0,1472	42/85
Meilleure évaluation	*0,3441*	*1/85*	*0,2739*	*1/85*	*0,2265*	*1/85*

Tableau 2.7 – Évaluations INEX 2006 pour la tâche Focused suivant la quantification generalised

Les résultats pour la sous-tâche *Focused* restent du même ordre que l'année précédente pour la quantification *generalised*. Ceci renforce l'idée que le changement de collection ne semble pas avoir eu d'impact sur les performances de notre SRI.

Par ailleurs, l'utilisation de la formulation de requête avec contraintes de structure dégrade les résultats. La sous-tâche *Focused* s'intéresse à un sous-ensemble restreint d'éléments XML l'application de contraintes sur la structure élimine des éléments XML pourtant jugés pertinents par

l'utilisateur. Il semble donc que l'ajout d'indications de structure n'apporte pas d'information profitable dans la description du besoin d'information.

Pour la sous-tâche *BestInContext*, un traitement « post-recherche » a été ajouté pour filtrer le résultat. En cas de chevauchement entre des éléments constituant le résultat, l'élément possédant le meilleur score est conservé.

BestInContext	BEP-D					
Soumission («Run»)	At 0.01	Rang	At 1.0	Rang	At 100.0	Rang
Run1CO2006. BestInContext	0,1175	30/77	0,2958	30/77	0,6430	39/77
Run2CO2006. BestInContext	0,1088	35/77	0,2907	34/77	0,6711	30/77
Meilleure évaluation	*0,1963*		*0,4089*		*0,7999*	

Tableau 2.8 – Évaluations INEX 2006 pour la tâche BestInContext

Les résultats pour cette nouvelle sous-tâche (tableau 2.8) restent du même ordre que pour les autres sous-tâches. Notre SRI peut donc être étendu par des modules complémentaires pour répondre à des tâches de recherche différentes.

Par ailleurs, d'autres expérimentations ont montré qu'à l'instar de la sous-tâche *Focused* l'utilisation de la formulation de requête avec contraintes de structure dégrade les résultats. La sous-tâche *BestInContext* s'intéresse également à un sous-ensemble restreint d'éléments XML. Ceci semble confirmer que l'ajout d'indications de structure n'apporte pas d'information profitable dans la description du besoin d'information lorsque la recherche se limite à des éléments spécifiques.

4 Bilan et perspectives

Dans ce chapitre, nous avons décrits nos contributions concernant la prise en compte de la structure des documents dans le processus de RI. L'objectif a été de prendre en compte les différents aspects introduits par l'utilisation de la structure des documents tant du point de vue des documents que de l'expression du besoin d'information. En effet, dans le cadre de la RI XML, l'utilisateur peut spécifier la granularité qu'il souhaite pour la réponse et également indiquer des contraintes sur la localisation des concepts qu'il recherche dans la structure sur les éléments à restituer.

4.1 Contributions

Pour répondre à la problématique soulevée par la RI XML, notre stratégie a été d'étendre le modèle de recherche précédemment développé pour une recherche non structurée (cf. chapitre 1) sans remettre en cause les contributions précédentes. Nos propositions se situent à différents niveaux : au niveau de l'indexation des documents et des requêtes, et au niveau de l'appariement. Au niveau de l'indexation des documents, nous avons considéré comme unité d'indexation les éléments XML possédant un contenu textuel. L'hypothèse a été que l'indexation au niveau le plus

fin (représenté par les éléments ayant un contenu textuel) offre le plus de possibilités de traitements ultérieurs. Au niveau de l'indexation des requêtes, le principe est d'associer chaque concept recherché à sa localisation souhaitée par l'utilisateur, et d'inclure dans la représentation de la requête la granularité souhaitée pour les éléments restitués. Au niveau de l'appariement, pour les éléments XML possédant un contenu textuel, nous avons étendu le principe d'appariement en intégrant une fonction qui estime la correspondance entre les localisations de concepts et une fonction qui estime la correspondance entre les granularités des éléments. Pour prendre en compte l'organisation hiérarchique des éléments XML, c'est-à-dire la présence d'éléments XML composés d'autres éléments XML, un principe d'agrégation de score a été défini.

Nos participations successives aux campagnes d'évaluation INEX de 2003 à 2006 ont permis de mener différentes expérimentations et d'évaluer l'efficacité des différentes propositions et leurs limites. Elles ont également renforcé mes connaissances en évaluation de SRI, initiées lors d'une participation à TREC (Benammar *et al.*, 2002a), qui ont pu être mises à profit dans d'autres travaux (Cabanac *et al.*, 2010b,a, 2011a; Palacio *et al.*, 2010b,a; Cabanac *et al.*, 2011b).

Les propositions ont été mises en œuvre, pour partie, dans le cadre du projet européen Ws-Talk et du projet Égide Alliance Grande-Bretagne QUEST.

4.2 Encadrement et diffusion scientifique

Les travaux menés dans cet axe ont donné lieu à différentes publications :

Thème	Publications
RI XML	Workshops internationaux : – INEX'06 (Hubert, 2007) – INEX'05 (Hubert, 2006) – INEX'04 (Hubert, 2005) – INEX'03 (Pinel-Sauvagnat *et al.*, 2003) Conférence nationale : – CORIA'05 (Hubert *et al.*, 2005)
Manipulation multidimensionnelle	Chapitre d'ouvrage international : – AKDM (Hubert et Teste, 2010) Conférence nationale : – EGC'09 (Hubert et Teste, 2009)

Une partie de ces travaux a permis le déroulement d'un stage de Master Recherche d'un étudiant en Master Recherche, Moussa Baraou Hamza (2007), dont j'ai assuré l'encadrement.

4.3 Perspectives

Les perspectives à ces travaux visent, d'une part, à faire évoluer le modèle pour traiter les tâches identifiées dans INEX depuis nos participations comme la recherche de passages plutôt que d'éléments XML. En effet, notre approche permet actuellement de restituer des composants d'un document XML. S'il s'avère que la meilleure réponse pour un utilisateur ne se situe pas au niveau

des éléments mais plutôt au niveau de passages, il est nécessaire d'étudier la manière d'étendre notre modèle pour prendre en compte cet aspect par exemple en travaillant sur la position des termes qui contribuent au score d'un élément XML. Par ailleurs, la tâche relative au retour de pertinence introduite récemment dans INEX peut constituer un cadre intéressant pour mener des travaux dans le domaine de la RI XML en lien avec nos contributions sur la prise en compte de l'utilisateur dans la RI (cf. chapitre 3).

D'autre part, la difficulté de manipulation des résultats de recherche obtenus soulève une nouvelle problématique. Une approche que nous souhaitons mener est une manipulation multidimensionnelle. Ce type d'approche s'appuiera sur les travaux déjà réalisés dans notre groupe dans le contexte d'une recherche d'information basée sur des hiérarchies de concepts au travers du prototype DocCube (Mothe *et al.*, 2003) et et de la plateforme d'analyse de données Tétralogie [3] (Hubert *et al.*, 2001) et dans notre équipe sur l'analyse OLAP et les entrepôts de documents XML (Khrouf et Soulé-Dupuy, 2005; Ravat *et al.*, 2010). Elle s'appuiera également sur les travaux d'extension de modèle multidimensionnel et de langage de manipulation OLAP que nous avons initié (Hubert et Teste, 2009, 2010). En effet, les modèles multidimensionnels permettent habituellement d'explorer les données suivant des niveaux d'analyse préétablis. Si nous considérons l'organisation hiérarchique d'un document comme la définition de niveaux d'analyse de tels modèles permettent d'explorer les données niveau par niveau. Notre modèle multidimensionnel propose une manipulation de données issues de différents niveaux de granularité permettant une exploration multidimensionnelle des résultats de RI XML.

3. `http://atlas.irit.fr`

3 Considérer l'utilisateur lors de la RI

1 Introduction

À l'origine, et encore à l'heure actuelle, les propositions de SRI ont eu pour objectif d'améliorer les performances de recherche, principalement du point de vue de la précision du résultat, par rapport aux requêtes soumises indépendamment des utilisateurs. Ainsi, la plupart des SRI traitent les requêtes de manière identique quel que soit l'utilisateur à l'origine de la requête. Pour la même requête le résultat est donc identique quel que soit l'utilisateur. Pourtant, pour le même besoin exprimé, l'appréciation du résultat est variable d'un utilisateur à l'autre et même pour un utilisateur d'une recherche à l'autre. Il semble donc important de considérer l'utilisateur dans le processus de recherche, ceci afin de fournir à chaque utilisateur un résultat qui le satisfasse.

Ce chapitre présente tout d'abord, dans la section 2, les propositions de la littérature qui prennent en compte l'utilisateur dans le processus RI dans des collections de documents. La section 2 introduit également les orientations de nos travaux. Nos propositions pour tenir compte de l'utilisateur dans la RI sont ensuite détaillées dans la section 3. La section 4 fait le bilan de nos contributions et présente les perspectives de travaux prévues.

2 Problématique et travaux de la littérature

Il est possible de considérer l'utilisateur à différents niveaux dans le processus de recherche d'information :

- dans l'expression du besoin utilisateur,
- dans la représentation des documents,
- dans l'appariement entre document et requête,
- dans la restitution du résultat à l'utilisateur.

Au niveau de l'expression du besoin, il est possible par exemple :

- d'utiliser des informations liées à l'utilisateur comme ses centres d'intérêts pour enrichir sa requête,
- ou encore d'utiliser des informations liées à d'autres utilisateurs comme les termes employés dans leurs requêtes pour aider un nouvel utilisateur dans la construction de sa requête.

Au niveau de l'indexation des documents, il est possible par exemple :

- d'étendre la représentation des documents par les termes de requêtes qui ont conduit à la restitution de ces documents par le SRI,
- d'étendre la représentation des documents par des termes issus d'annotations ou « tags » posées par les utilisateurs.

Vis-à-vis de l'appariement entre documents et requêtes, il est possible par exemple de définir des similarités en fonction des informations liées à l'utilisateur.

Enfin, pour la restitution du résultat il est possible par exemple :

- de présenter le résultat en tenant compte de caractéristiques de l'utilisateur comme une déficience visuelle,
- d'organiser les documents du résultat suivant des préférences de l'utilisateur.

Ces différents aspects impliquent autant de problématiques présentées dans les sections suivantes en lien avec les propositions de la littérature.

2.1 Sources d'informations sur l'utilisateur

Une problématique concerne la collecte ou l'extraction d'informations sur l'utilisateur. Un premier type d'approches consiste à obtenir des informations de l'utilisateur sur les documents restitués par le SRI. Ces informations collectées sont les jugements de pertinence sur un ensemble de documents extraits du résultat fourni par une première recherche. Ces informations peuvent être demandées explicitement à l'utilisateur qui indique parmi les documents présentés ceux qu'il juge pertinents et ceux qu'il juge non pertinents (Salton et Buckley, 1990; Belkin *et al.*, 1996; Iwayama, 2000; Rooney *et al.*, 2006). Dans le but de ne pas solliciter l'utilisateur, elles peuvent également être estimées implicitement en observant les actions de l'utilisateur sur les documents comme le temps de lecture (Kelly et Belkin, 2001), le passage de la souris sur les documents (White *et al.*, 2002) ou les clics (Joachims, 2002; Jung *et al.*, 2007).

Outre la conservation explicite d'informations relatives aux utilisateurs que sont les profils utilisateurs, les fichiers de connexion (« logs ») constituent également une forme de conservation d'informations indirectes sur les utilisateurs en identifiant des relations entre termes des requêtes et termes des documents (Cui *et al.*, 2003b; Zhu et Gruenwald, 2005), ou en classifiant les requêtes en fonction des clics sur les documents restitués (Baeza-Yates *et al.*, 2004). Des informations extraites de fichiers de connexions peuvent notamment être utilisées pour construire des profils utilisateurs (Martín-Bautista *et al.*, 2002). Carman *et al.* (2008) utilisent quant à eux les annotations (« tags ») et l'historique des signets (« bookmark history ») pour construire des profils. Psarras et Jose (2006) surveillent l'usage du système par l'utilisateur, c'est-à-dire les documents regardés et autres actions de l'utilisateur, pour construire un profil et inférer les changements d'intérêts de l'utilisateur.

2.2 Représentation des informations sur l'utilisateur

L'utilisation d'informations liées aux utilisateurs implique généralement leur conservation. Une problématique est donc d'identifier les informations relatives aux utilisateurs qu'il est possible et utile de conserver et de définir des modèles pour les représenter et les stocker.

L'identification des informations relatives aux utilisateurs qu'il est possible et utile de conserver à fait l'objet de nombreux travaux principalement au travers de la notion de profil utilisateur. Un profil utilisateur regroupe généralement des informations personnelles relatives à un utilisateur et des informations relatives à ses centres d'intérêt. Pretschner et Gauch (1999) conservent, par exemple, certaines caractéristiques des pages Web visitées par l'utilisateur tels que leurs catégories de sujets, leur taille et le temps passé sur la page. Martín-Bautista *et al.* (2002) définissent le profil utilisateur à partir d'informations sur le comportement de navigation et sur l'environnement personnel de l'utilisateur. Danilowicz et Nguyen (2002) conservent un ensemble de mots des documents consultés par l'utilisateur ; Arezki *et al.* (2004) y ajoutent en plus que les fréquences des mots dans les documents et les fréquences de cooccurrence entre termes. Gils et Schabell (2003) définissent un profil comme un ensemble de préférences sur le comportement du SRI (par exemple, le nombre de résultats par page) et de contraintes sur les résultats (par exemple, le format des documents du résultat). Pour Lainé-Cruzel (1999), un profil utilisateur est vu comme un regroupement d'informations sur qui est l'utilisateur (par exemple, niveau de connaissances dans le domaine), sur ce qu'il veut (par exemple, volume d'informations attendu) et sur ce qu'il fera de l'information (par exemple, tâche qui implique sa recherche). Amato et Straccia (1999) soulignent qu'au travers du profil utilisateur il faut décrire le « quoi » c'est-à-dire ce qui doit être représenté et le « comment » c'est-à-dire comment le représenter. Cinq catégories d'informations différentes sont distinguées : les données personnelles, les données de collecte qui concernent les préférences et restrictions sur les documents recherchés, les données de distribution qui concernent les préférences sur les modes de distribution des informations collectées, les données d'actions et les données de sécurité. Kostadinov (2007) identifie quant à lui cinq principales dimensions pour un profil utilisateur : le domaine d'intérêt, les données personnelles, les données de qualité, les données de livraison et les données de sécurité. Il complète la notion de profil utilisateur par celles de contexte qui décrit l'environnement d'interaction entre utilisateur et système et de préférence pour définir l'importance des informations d'un profil ou d'un contexte. Dans une optique d'intégration, Chevalier *et al.* (2005) proposent un modèle générique de profil qui permet de décrire tout type de profil pour l'accès à des ressources et permet également l'interopérabilité entre différents profils.

Les profils utilisateurs sont principalement représentés sous forme de vecteurs de mots-clés (Göker et McCluskey, 1991; Armstrong *et al.*, 1995; Danilowicz et Nguyen, 2002; Jeon *et al.*, 2008). D'autres représentations plus riches ont été proposées comme des réseaux sémantiques (Stefani et Strappavara, 1998; Arezki *et al.*, 2004), des hiérarchies de connaissances (Berisha-Bohé et Rumpler, 2007), des hiérarchies de concepts ou d'ontologies (Pretschner et Gauch, 1999; Nanas *et al.*, 2003; Sieg *et al.*, 2007).

D'autres travaux ont également cherché à conserver les requêtes passées des utilisateurs et les résultats obtenus directement (Raghavan et Sever, 1995; Selberg et Etzioni, 1998; Klink, 2004), ou sous formes de matrices (Balfe et Smyth, 2005), de séries temporelles (Trousse *et al.*, 1999) ou encore de cas pour des systèmes de raisonnement par cas (Jéribi et Rumpler, 2002; Iszlai et Egyed-Zsigmond, 2006).

2.3 Exploitation des informations sur l'utilisateur

L'exploitation des informations relatives aux utilisateurs et à leur utilisation des SRI a fait l'objet de différentes propositions notamment dans :

- la reformulation de requêtes à partir des requêtes soumises précédemment et conservées par les utilisateurs (Raghavan et Sever, 1995; Fitzpatrick et Dent, 1997; Klink, 2004), à partir de fichier de logs (Cui *et al.*, 2003b; Zhu et Gruenwald, 2005), de profils d'utilisateurs (Arezki *et al.*, 2004; Kostadinov, 2007; Zayani *et al.*, 2009), sur les documents stockées par l'utilisateur (Calegari et Pasi, 2008), ou sur les « tags » posés par les utilisateurs (Carman *et al.*, 2008),
- la recommandation de requêtes à partir de fichier de logs (Baeza-Yates *et al.*, 2004; Shi et Yang, 2007; Liu *et al.*, 2008),
- la recommandation de documents à partir des requêtes soumises précédemment par les utilisateurs et conservées avec leurs résultats (Raghavan et Sever, 1995),
- l'extension de l'indexation des documents à partir des requêtes soumises (Selberg et Etzioni, 1998; Kemp et Ramamohanarao, 2002; Amitay *et al.*, 2005),
- l'extension de la fonction d'appariement entre requêtes et documents en intégrant les informations des profils utilisateurs (Speretta et Gauch, 2005; Tamine *et al.*, 2007; Naderi *et al.*, 2007; Mylonas *et al.*, 2008; Hadjouni *et al.*, 2009), en intégrant les requêtes précédemment soumises (Balfe et Smyth, 2005), en utilisant des fichiers de logs (Zhuang et Cucerzan, 2006),
- la recherche d'utilisateurs similaires en fonction des profils des utilisateurs (Denos *et al.*, 2004; Jeon *et al.*, 2008; Naderi *et al.*, 2008) ou de l'usage des documents (Cabanac, 2008).

2.4 Orientations de nos contributions

Les propositions qui visent à prendre en compte l'utilisateur dans le processus de recherche d'information se sont principalement intéressés à la représentation de l'utilisateur notamment du point de vue de ses centres d'intérêt ainsi qu'à l'exploitation des informations utilisateurs pour la reformulation de requête ou l'appariement entre requêtes et documents.

Une problématique importante reste la gestion de l'évolution des informations relatives aux utilisateurs tant au niveau de la conservation des informations que de leur exploitation. La gestion de l'évolution des informations relatives aux utilisateurs concerne essentiellement les approches basées sur des profils utilisateurs vis-à-vis de la mise à jour de ces derniers. Les approches fondées sur les fichiers de connexion et le stockage des requêtes passées doivent quant à elles gérer l'ajout constant d'informations relatives aux nouvelles recherches effectuées par les utilisateurs. La mise à jour de profils d'utilisateurs est réalisée :

- de manière explicite, à partir des retours de pertinence exprimés, par les utilisateurs sur les résultats obtenus (Göker et McCluskey, 1991; Danilowicz et Nguyen, 2002; Mylonas *et al.*, 2008),
- ou de manière implicite, en fonction des actions effectuées par les utilisateurs par rapport aux résultats tels que l'ouverture ou non d'un document ou le temps passé sur un document (Pretschner et Gauch, 1999; Chen et Zeng, 2008).

Peu de travaux ont abordé le problème de l'évolution au regard du nombre de travaux basés sur l'utilisation d'informations liées aux utilisateurs. C'est vers cette problématique que se sont orientés nos travaux, d'une part, autour d'une approche basée sur des profils d'interrogation et

leur évolution, d'autre part, autour d'une approche basée sur la réutilisation d'expériences de recherche reposant sur des principes de gestion de versions. La problématique de l'exploitation des informations dans le cadre d'un processus de recherche est également une de nos préoccupations. Une proposition consiste à utiliser les requêtes précédemment posées par les utilisateurs pour construire un graphe permettant la définition progressive d'une nouvelle requête par un utilisateur.

3 Recherche exploitant des profils

La recherche à base de profils regroupe plusieurs aspects. Dans nos travaux, la notion de profil correspond à la conservation d'un ensemble d'informations qui est ensuite utilisé pour améliorer le processus de recherche ou adapter la restitution du résultat. Nous nous sommes intéressés dans un premier temps à l'utilisation des profils pour améliorer le processus de recherche. Dans ce cadre, l'utilisation de profil rassemble trois problématiques différentes :

- un premier aspect concerne les informations à conserver dans le but d'être exploitées lors des recherches. Il s'agit de définir et de structurer les informations qui seront stockées.
- un deuxième aspect concerne l'exploitation des profils c'est-à-dire comment utiliser les profils lors des recherches d'information.
- enfin un dernier aspect concerne la mise à jour des profils : comment faire évoluer les profils de manière à maintenir leur représentativité et leurs possibilités d'exploitation ? Cet aspect est resté peu étudié dans la littérature bien qu'il représente un enjeu important.

Nous nous sommes intéressés principalement au dernier aspect qui concerne l'évolution des profils. Cet aspect est crucial si l'on veut maintenir une utilisation pertinente, efficace et durable des profils. La gestion de l'évolution nécessite bien sûr de prendre en compte les deux premiers aspects c'est-à-dire la définition de profils et leur exploitation.

Nous avons proposé deux approches qui seront développées dans les sections suivantes :

- l'approche basée sur les profils d'interrogation et leur mise à jour qui s'appuie sur des principes de reformulation de requêtes (Benammar *et al.*, 2001, 2002b, 2003),
- l'approche basée sur la réutilisation d'expériences de recherche qui s'appuie sur des principes de gestion de versions (Hubert et Mothe, 2007b).

3.1 Profils d'interrogation

Les profils d'interrogation s'attachent s'attache à décrire l'utilisateur au travers des recherches d'information qu'il effectue. De nombreux travaux se sont intéressés à la description de l'utilisateur notamment au travers d'informations propres à l'individu c'est-à-dire des caractéristiques personnelles et à ces centres d'intérêt. Nous nous sommes focalisés sur l'identification des centres d'intérêt au travers des différentes interrogations réalisées par l'utilisateur (Benammar *et al.*, 2001). Ces travaux peuvent naturellement être complétés par la gestion de profil d'utilisateur du point de vue des caractéristiques personnelles. Ces travaux ont été menés dans le cadre de la thèse d'Anis Benammar (2003), co-encadrée avec le professeur Josiane Mothe.

3.1.1 Modélisation de profils d'interrogation

Un profil d'interrogation correspond à la représentation d'un centre d'intérêt ou d'un besoin d'information. Notre approche considère qu'un centre d'intérêt rassemble un ensemble de besoins d'informations exprimés au travers de différentes interrogations ou requêtes. Cette considération est traduite à travers deux niveaux de profils d'interrogation (Benammar *et al.*, 2003) :

- un profil d'interrogation à court terme représente une session de recherche. Ce type de profil correspond à un besoin d'information à un instant donné.
- un profil d'interrogation à long terme est construit à partir des différentes sessions de recherche d'un utilisateur suivant le même besoin d'information. Ce type de profil correspond plus à un centre d'intérêt de l'utilisateur c'est-à-dire qui perdure.

À chaque fin de session de recherche, le profil d'interrogation à court terme devient un profil à long terme qui pourra être réutilisé.

Initialement, un profil d'interrogation à court terme correspond à une requête utilisateur. Alternativement, le profil d'interrogation à court terme peut correspondre à un profil d'interrogation à long terme existant. Dans ce cas, l'utilisateur choisit de repartir d'un profil existant pour démarrer une nouvelle session de recherche. La figure 3.1 illustre la définition des deux niveaux de profils.

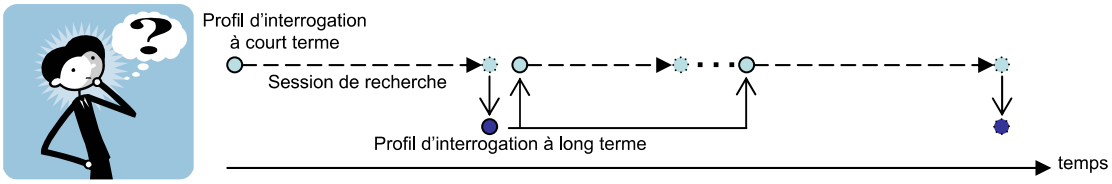


Figure 3.1 – Profils d'interrogation à court terme et à long terme

La validation d'un profil d'interrogation à court terme implique sa transformation en profil à long terme. Lorsque le profil d'interrogation à court terme correspond à une nouvelle requête la validation crée un nouveau profil à long terme. Lorsque le profil à court terme est issu d'un profil à long terme, la validation du profil à court terme correspond à la mise à jour du profil à long terme dont il est issu.

L'évolution du profil d'interrogation à court terme peut être réalisée explicitement par l'utilisateur qui modifie la requête en cours en ajoutant ou supprimant des termes. Elle peut également être réalisée de manière automatique notamment en exploitant les documents obtenus en résultat, particulièrement le résultat obtenu avant validation du profil. L'évolution du profil d'interrogation réside donc en une succession de reformulations appliquées automatiquement ou proposées à l'utilisateur sous forme de recommandations. Nous avons rapproché la reformulation de profil d'interrogation de la reformulation de requête. Nous nous sommes donc intéressés aux techniques de reformulation automatique de requête.

3.1.2 Reformulation automatique de profil d'interrogation

La reformulation de profil s'appuie sur les travaux relatifs à la reformulation de requête. L'idée est de rassembler les informations liées à un centre d'intérêt ou un besoin d'information afin de le

décrire au mieux.

Différentes méthodes de reformulation de requête ont été proposées faisant intervenir l'utilisateur ou non. La reformulation automatique de requête peut rassembler plusieurs aspects comme l'expansion de la requête par de nouveaux termes issus de documents pertinents (Rocchio, 1971; Buckley *et al.*, 1994; Eguchi, 2000) ou la repondération des termes de la requête étendue (Rocchio, 1971).

Pour notre approche de reformulation de profil, une solution serait d'impliquer l'utilisateur en lui demandant de préciser dans le résultat de recherche, les documents qu'il juge pertinents. Une autre approche serait de se baser sur le résultat de recherche lorsqu'il valide la recherche, en faisant l'hypothèse qu'il valide celle-ci uniquement lorsque le résultat correspondant est satisfaisant pour lui. Cependant dans ce cas les documents considérés pertinents par l'utilisateur ne sont pas identifiés dans le résultat. Néanmoins, dans le but d'éviter d'impliquer l'utilisateur dans la phase de reformulation du profil, nous avons opté pour la seconde solution. Pour cela nous nous sommes appuyés sur les méthodes qui utilisent les documents les mieux classés dans les résultats de recherche. Ce type d'approche est appelé pseudo-retour de pertinence (« blind feedback ») (Xu et Croft, 2000; Eguchi, 2000) puisque la pertinence n'est pas donnée par l'utilisateur. Les approches usuelles de pseudo-retour de pertinence utilisent les documents les mieux classés pour la requête initiale pour créer une nouvelle requête (Kwok *et al.*, 1997; Baeza-Yates et Ribeiro-Neto, 1999). L'amélioration des résultats dépend de la qualité des documents les mieux classés ; des expérimentations menées dans le cadre de TREC ont montré l'irrégularité des performances des techniques qui utilisent les premiers documents (Xu et Croft, 1996). Lorsque des documents non pertinents existent parmi les mieux classés, la reformulation peut s'avérer inefficace.

Notre approche s'est donc concentrée sur la qualité de l'ensemble de documents utilisé pour la reformulation d'un profil (Benammar *et al.*, 2002b). La figure 3.2 illustre les différentes étapes du processus de reformulation automatique proposé.

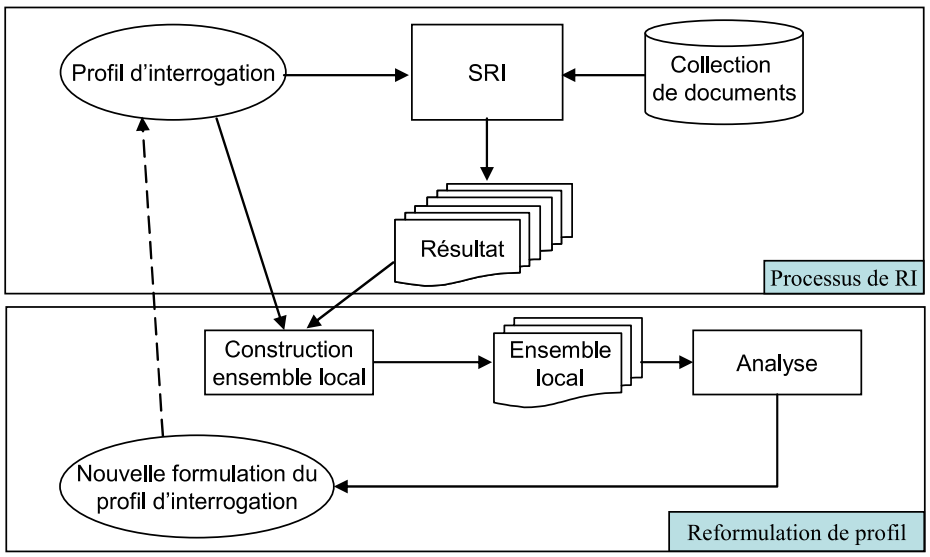


Figure 3.2 – Processus de reformulation automatique (Benammar *et al.*, 2002b)

Un processus traditionnel de recherche d'information est tout d'abord appliqué pour retrouver un ensemble de documents en réponse au profil de requête à court terme. Le processus de

reformulation utilise ensuite une approche alternative de retour local de pertinence. La principale différence réside dans l'application d'une première étape qui consiste à construire un ensemble local de documents à partir du résultat de recherche. L'objectif est donc d'améliorer la qualité de l'ensemble de documents utilisé pour la modification du profil. Cette étape supplémentaire vise à réordonner les documents initialement retrouvés dans le but de repositionner en tête de liste des documents pertinents initialement moins bien classés. Les premiers documents sont réordonnés en fonction du profil en cours en appliquant une mesure de similarité (différente de la recherche initiale) entre les documents et le profil. Les nouveaux documents les mieux classés, à l'issue de cette étape, constituent l'ensemble local de documents sur lequel est appliquée l'analyse de cooccurrence. Le but de cette analyse est de mesurer les liens qui existent entre les termes des documents et le profil. Les termes, issus des documents de l'ensemble local, ayant les plus forts liens sont ensuite utilisés pour étendre le profil. Les liens entre termes et profil sont calculées à partir des cooccurrences de termes comme suit :

$$lien(t_i, P) = \sum_{j \in P} cooccurrence(t_i, p_j) \tag{3.1}$$

où t_i est un terme des documents de l'ensemble local et p_j un terme du profil d'interrogation P.

Le degré de cooccurrence entre termes est calculé suivant la définition de (Xu et Croft, 2000) :

$$cooccurrence\left(t_i, p_j\right) = \log_{10}\left(1 + \sum_{d \in S} tf\left(t_i, d\right) \cdot tf\left(p_j, d\right)\right) \cdot idf\left(t_i\right) / \log_{10}\left(n\right) \tag{3.2}$$

où $tf_{(ti,d)}$ et $tf_{(p_j,d)}$ sont les fréquences des termes t_i et p_j dans le document d,
n est le nombre de documents dans l'ensemble local de documents,
$idf\left(t_i\right) = \log_{10}\left(\frac{N}{N_i}\right)$,
N est le nombre de documents dans la collection,
N_i est le nombre de documents de la collection contenant le terme t_i.

3.1.3 Expérimentations sur le corpus Ohsumed

L'objectif des expérimentations outre la validation de l'approche proposée a été de mesurer l'apport d'un tel mécanisme dans la gestion de l'évolution des profils. De plus les expérimentations ont permis d'évaluer l'apport de chaque étape dans le processus de reformulation.

Ces expérimentations ont été menées sur le jeu de données d'apprentissage de la tâche Filtrage de la campagne d'évaluation TREC-9 déjà utilisé pour d'autres expérimentations décrites précédemment (cf. chapitre 1, section 3.2.3). Pour compléter les caractéristiques données sur la collection de documents OHSUMED-87 utilisée, celle-ci comporte 46 095 termes distincts et en moyenne 130 termes par document.

Nous nous sommes basés sur les 63 requêtes du jeu d'essai pour simuler des profils utilisateurs à court terme.

Nous avons utilisé le programme trec_eval [1] pour calculer les valeurs de précision à 10, et 20 documents restitués ainsi que les valeurs de précision moyenne (MAP).

1. http://trec.nist.gov

Différents paramètres ont été fixés pour réaliser nos expérimentations :

- le nombre de documents utilisés pour construire l'ensemble local fixé à 200 ; l'ensemble local est donc construit à partir des 200 premiers documents du résultat initial,
- la taille de l'ensemble local est fixée à 5,
- le nombre de termes ajoutés à l'expression du profil ; le nombre de termes à ajouter est fonction du nombre de documents de l'ensemble local. Nous nous sommes basés sur la formule définie dans (Belkin *et al.*, 1999), Nbtermes = $(5 \times m) + 5$, où m est le nombre de documents de l'ensemble local,
- la mesure cosinus utilisée pour définir les similarités entre documents.

Le principe des expérimentations a été de comparer les évaluations des résultats obtenus :

- par la recherche initiale,
- en utilisant directement les premiers documents de la recherche initiale pour reformuler le profil,
- en appliquant notre processus complet de reformulation du profil.

Sur les 63 profils d'interrogation utilisés, les résultats se sont avérés invariants entre recherche initiale et recherche après processus complet de reformulation pour 25 d'entre eux (39,7 %). Après analyse, deux cas apparaissent :

- Pour 15 profils, la recherche initiale n'a restitué aucun document pertinent, donc aucun dans les 200 premiers documents (ceux utilisés pour construire l'ensemble local). Dans ce cas, l'approche de reformulation ne possède aucun élément pour améliorer le profil.
- Pour 10 profils, bien que le résultat initial contienne des documents pertinents le résultat demeure invariant. En fait, l'ensemble local construit reste vide, aucun document ne donnant une similarité suffisante avec le profil initial. Bien que ces cas ne soient pas pénalisants en termes de précision, il serait intéressant d'examiner d'autres mesures de similarité dans le but d'obtenir une amélioration plutôt qu'une invariance.

Parmi les 38 profils restants, le processus de reformulation améliore les évaluations pour 26 d'entre eux et dégrade les résultats pour 12 d'entre eux.

Le tableau 3.1 présente les résultats des comparaisons pour les 48 profils conduisant à des évaluations non nulles, suivant les mesures P10, P15, P20 et MAP :

	P10	P15	P20	MAP
Résultat initial (référence)	0,387	0,326	0,276	0,395
Résultat après reformulation sans construction d'ensemble local	0,427	0,342	0,290	0,417
Résultat après reformulation et construction d'ensemble local	0,427	0,361	0,305	0,419

Tableau 3.1 – Amélioration de l'efficacité du SRI grâce aux reformulations de profils

Les expérimentations montrent que le processus de reformulation améliore en moyenne (MAP) les évaluations (+ 6 %) pour la reformulation avec construction d'ensemble local par rapport au résultat initial). Elles montrent également que l'étape de reclassement des documents (construction de l'ensemble local) améliore légèrement la qualité des documents utilisés dans l'analyse puisque les résultats sont légèrement supérieurs (+ 5,5 % pour la reformulation sans construction d'ensemble local par rapport au résultat initial). Cependant, les tests de randomi-

sation effectués n'indiquent pas que la différence observée est significative. Une hypothèse expliquant la faible amélioration réside sans doute dans le fait que le comportement de la fonction de reclassement utilisée n'est pas suffisamment différent de celui de la fonction d'appariement initiale. Suivant les principes de combinaison de résultats définis dans (Fox et Shaw, 1994) et (Lee, 1997), nous avons étudié la combinaison de différentes méthodes de reclassement pour construire l'ensemble local. Ces expérimentations complémentaires n'ont cependant pas montré d'amélioration par rapport à l'application d'une méthode unique.

D'autres études doivent donc être menées pour améliorer les résultats plus particulièrement du point de vue des 19 % de cas où l'on observe une dégradation et notamment explorer de nouvelles méthodes de construction de l'ensemble local.

Par ailleurs, plutôt que de synthétiser les évolutions de profil au travers de modification du contenu du profil, une seconde approche que nous menons consiste à conserver les évolutions au travers d'un mécanisme de gestion de versions (Hubert et Mothe, 2007b).

3.2 Réutilisation d'expériences de RI

Les différentes interrogations soumises au SRI correspondant à un même besoin d'information peuvent permettre de faire évoluer un profil d'interrogation. Ce profil peut ensuite être réutilisé pour des recherches suivantes. La réutilisation peut consister à repartir du profil conservé, proposer les profils en rapport avec une nouvelle requête, ou proposer des termes de ces profils pour guider l'utilisateur. Cependant, cette stratégie permet d'exploiter uniquement les termes utilisés dans les recherches correspondant au même besoin d'information. La réutilisation d'expériences de recherche passées représente un moyen de guider plus finement l'utilisateur. L'idée directrice est d'exploiter les démarches d'interrogation de l'utilisateur ou d'un groupe d'utilisateurs, de manière à mettre une nouvelle interrogation en relation avec des interrogations déjà réalisées par le passé. Des suggestions issues des reformulations liées aux interrogations passées retrouvées permettrons ensuite d'aider l'utilisateur à poursuivre sa recherche d'information. Le point de départ de notre approche est la notion de session de reformulation de requête (Amitay *et al.*, 2005). Ces travaux ont permis à une étudiante en Master Recherche Randa Al Sabbagh (2005), de réaliser son stage de recherche dont j'ai assuré l'encadrement.

3.2.1 Session de reformulations

Amitay *et al.* (2005) définissent une session de reformulations comme une série de reformulations dérivées d'une requête initiale dont l'objectif est de satisfaire un même besoin d'information. Ils considèrent, de plus, que la dernière reformulation de la requête conduit à un résultat satisfaisant l'utilisateur.

Une session de reformulations regroupe donc une succession de requêtes liées par des liens implicites. Cependant, ces liens ne sont habituellement pas conservés et par conséquent ne sont pas exploités, comme illustré dans la figure 3.3.

Dans notre approche, nous étendons le concept de session de reformulations aux résultats de recherche. Une session de reformulations regroupe donc une succession de couples (requête, résultats de recherche) liés par des liens d'évolution.

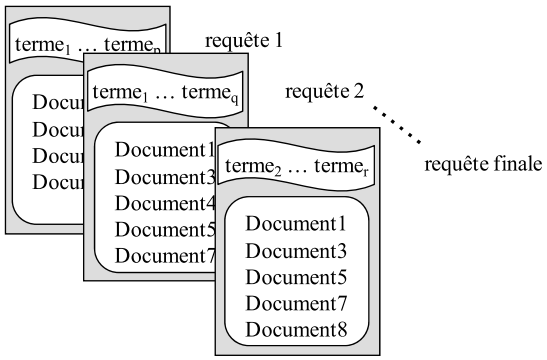


Figure 3.3 – Session de reformulation

3.2.2 Versions d'interrogations

Nos travaux s'intéressent aux liens implicites existant entre les différentes tentatives d'interrogation effectuées par un utilisateur pour satisfaire un même besoin d'information, qui peut être un besoin récurrent. Ces liens qui ne sont habituellement pas conservés sont pourtant une source d'informations utile. Ainsi, nous considérons les différentes interrogations successives comme des évolutions d'une recherche ; elles peuvent donc être modélisées sous formes de versions. Dans notre approche, nous conservons ces liens qui sont modélisés explicitement sous formes de liens « d'évolution » entre versions d'interrogations. Une interrogation regroupe une formulation de requête et le résultat restitué par le SRI. La figure 3.4 illustre le principe de versions d'interrogation.

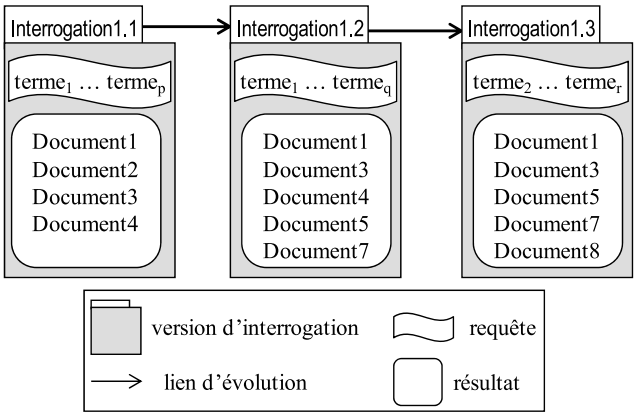


Figure 3.4 – Versions d'interrogations (Hubert et Mothe, 2007b)

3.2.3 Expérience de RI

Dans notre approche, la notion d'expérience de recherche étend l'idée de session de reformulation en cherchant à conserver toute la démarche d'un utilisateur qui utilise un SRI. L'idée est de conserver des informations relatives à chaque tentative d'interrogation menées par un utilisateur, c'est-à-dire de la requête soumise jusqu'à la restitution du résultat. L'idée est également de conserver l'évolution de l'interrogation menée par l'utilisateur. Une expérience de recherche est

donc modélisée comme un ensemble de versions d'interrogations liées par des liens d'évolutions (cf. figure 3.5).

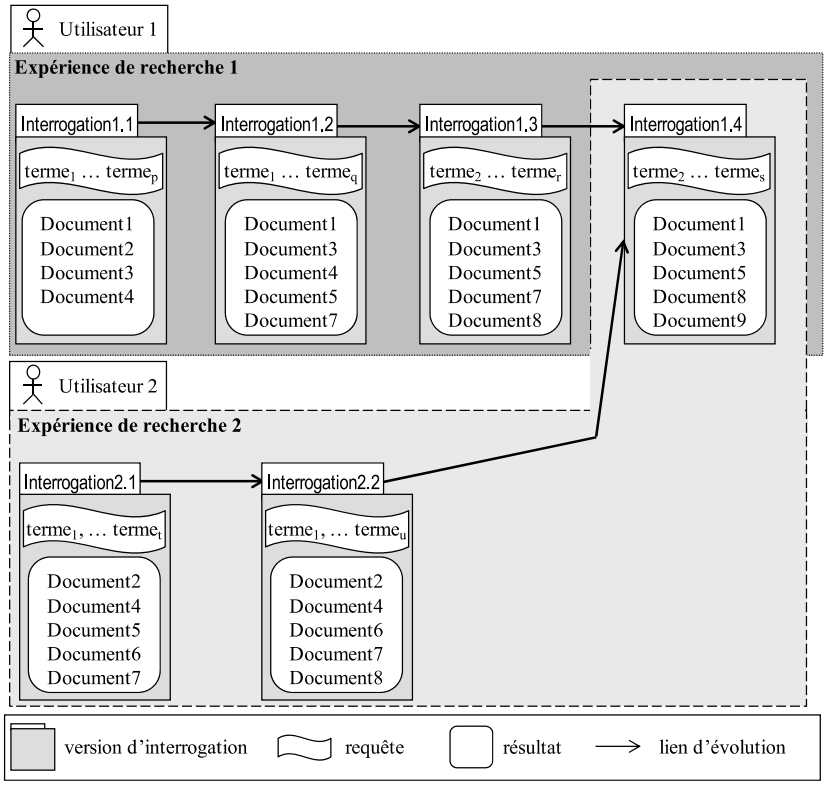


Figure 3.5 – Expériences de recherche (Hubert et Mothe, 2007b)

Un utilisateur rassemble plusieurs expériences de recherche. Plusieurs utilisateurs peuvent partager des expériences de recherche ou des parties d'expériences c'est-à-dire des interrogations. L'approche est de type « donnant-donnant » où un groupe bénéficie de l'expérience de chaque utilisateur et vice-et-versa.

3.2.4 Exploitation des versions d'interrogations

Notre approche au travers des versions d'interrogation permet différentes exploitations des recherches passées. Elle permet d'une part de proposer :

- des recommandations de requêtes pour remplacer la requête initiale,
- des recommandations de termes pour aider à la reformulation de requêtes,
- des recommandations de documents pour aider l'utilisateur dans l'exploitation des résultats.

L'exploitation des recherches passées conservées peut être réalisée lorsque l'utilisateur construit une requête. La requête est dans ce cas comparée aux versions de requêtes conservées Elle peut également intervenir après que l'utilisateur ait soumis une première requête au SRI et ait obtenu un premier résultat. Dans ce cas, la requête initiale et le résultat initial associé sont comparés aux versions d'interrogations conservées.

Si une forte similarité est estimée (par exemple supérieure à un seuil fixé par l'utilisateur ou déterminé après une phase d'apprentissage), différentes recommandations peuvent être faites à l'utilisateur :

- les mots-clés extraits des requêtes ou des documents issus des expériences similaires peuvent êtres utilisés pour des recommandations de termes,
- les documents jugés pertinents par d'autres utilisateurs dans les expériences similaires peuvent être proposés à l'utilisateur pour des recommandations de documents. Ces documents peuvent être ceux des résultats associés aux versions d'interrogation proches ou aux versions finales liées aux versions d'interrogation proches. Par exemple, dans la figure 3.5, si la version d'interrogation 1.2 est évaluée similaire à une nouvelle requête, la reformulation finale de la requête trouvée dans la version d'interrogation 1.4 liée à la version 1.2 peut être proposée à l'utilisateur,
- les reformulations finales de requêtes extraites des expériences similaires peuvent être suggérées à l'utilisateur dans le cadre de recommandations de requêtes,
- les documents jugés pertinents par d'autres utilisateurs dans les expériences similaires peuvent être proposés à l'utilisateur pour des recommandations de documents. Ces documents peuvent être ceux des résultats associés aux versions d'interrogation proches ou aux versions finales liées aux versions d'interrogation proches.

Par ailleurs, au-delà de la recherche d'expériences similaires à partir de version d'interrogation similaires, la gestion de versions pour les expériences de recherche permet également de rapprocher les utilisateurs en matière de recherche d'information soit suivant leurs centres d'intérêts soit suivant leur comportement de recherche. Il est possible de rapprocher les utilisateurs en évaluant :

- la proximité de leurs expériences de recherche respectives. Il s'agit dans ce cas d'étudier globalement les expériences de recherches de deux utilisateurs distincts,
- la similarité de deux expériences de recherche du point de vue de la progression de la recherche d'information. Il s'agit dans ce cas de comparer les versions d'interrogation en incluant leurs positions relatives.

La recherche d'expériences similaires et la comparaison d'expériences peuvent être basées sur différents critères :

- les requêtes,
- les résultats de recherche en tant que listes de documents,
- les documents des résultats en termes de contenu

La recherche d'expériences similaires et la comparaison d'expériences nécessitent l'utilisation voire la définition de mesures de similarité appropriées telles que :

- similarité entre requêtes,
- similarité entre listes de résultats,
- similarité entre documents,
- similarité entre individus fondée sur les précédentes.

3.2.5 Mise en œuvre

La mise en œuvre de notre approche comprend trois aspects :

– la modélisation des expériences de recherche,
– la gestion de versions d'interrogation,
– la recherche d'information intégrant les différents types de recherche et les mesures de similarité associées.

Du point de vue de la modélisation des expériences, notre approche considère :

– une expérience de recherche est un ensemble de versions d'interrogation liées par des liens d'évolution,
– une interrogation (ou version d'interrogation) est constituée d'une requête et d'un résultat,
– une requête est considérée comme une liste de mots-clés,
– un résultat est une liste de documents restitués par le SRI,
– un document est considéré comme un ensemble de mots-clés pondérés. Un document peut être jugé pertinent ou non par l'utilisateur, ou rester non jugé.

Les expériences de recherche peuvent donc être modélisées avec le diagramme de classes UML (représentation simplifiée) comme suit :

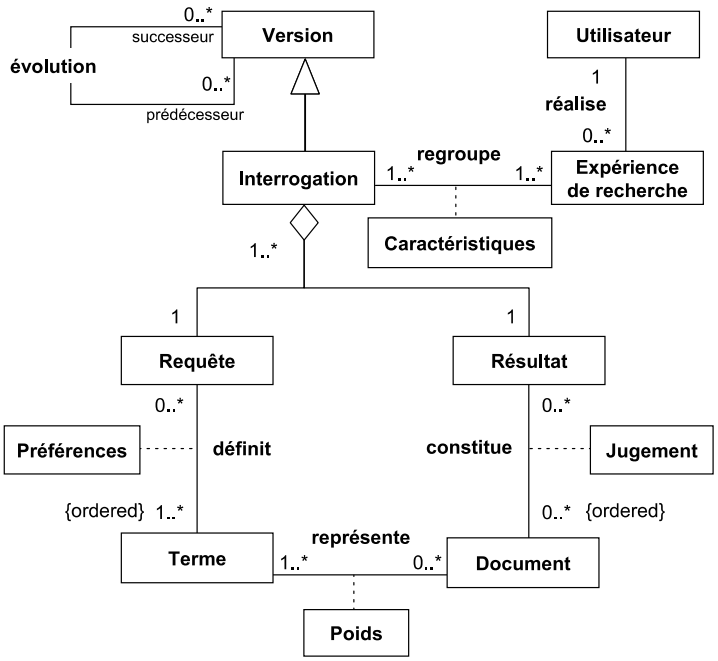


Figure 3.6 – Diagramme de classes UML décrivant les expériences de recherche(Hubert et Mothe, 2007b)

Du point de vue de la gestion de versions d'interrogations, différentes solutions sont possibles. De nombreux travaux se sont intéressés à la gestion de versions principalement dans le domaine du développement logiciel (Conradi et Westfechtel, 1998) et dans le domaine des bases de données (Katz, 1990; Andonoff *et al.*, 1998; Jomier et Cellary, 2000). Deux types d'approches sont considérées : la manipulation de versions au niveau d'un ensemble d'objets liés (Conradi et Westfechtel, 1998; Jomier et Cellary, 2000), ou la gestion de versions au niveau d'un objet (Katz, 1990; Andonoff *et al.*, 1998).

La gestion de versions d'interrogation proposée peut s'appuyer sur nos travaux antérieurs relatifs à la gestion de versions d'objets complexes. En effet, les interrogations telles que nous les

concevons sont des objets complexes. Dans ces travaux, nous proposons une approche pour la modélisation et la manipulation d'objets complexes dans les bases de données orientées objet (Hubert, 1997). Cette approche permet notamment la création de bases de données associant des classes pour lesquelles les versions d'objets sont gérées et des classes sans gestion de version. De plus, elle propose des langages (textuel de type SQL et graphique) permettant la manipulation de versions d'objets complexes.

Enfin, la recherche d'information intégrant les différents types de recherche et les mesures de similarité associées peuvent reposer sur le modèle générique précédemment décrit (cf. chapitre 1, section 3.1.2.2, équation 1.1) c'est-à-dire :

$$Score(BI_1, UI_1) = \left(\sum_{c \in C_1} imp_{BI}(c, BI_1) \cdot imp_{UI}(c, UI_1) \right) \cdot recouv(BI_1, UI_1)$$

Dans le cadre de la recherche d'information qui consiste à retrouver des documents textuels répondant à une requête définie sous forme de mots-clés, le SRI repose tout d'abord sur une mesure de similarité entre un document et une requête détaillée précédemment pour la recherche*ad hoc* par mots-clés (cf. chapitre 1, section 3.2.1). Le même calcul de score peut être utilisé pour estimer la similarité entre documents. En effet, dans le cas où l'on recherche des documents similaires à un document donné, ce dernier peut être assimilé à une requête.

De plus, la même base de SRI peut être adaptée pour évaluer les similarités entre requêtes et les similarités entre listes de résultats. L'adaptation intervient alors au niveau de la fonction de calcul de score appliquée compte tenu du fait que les éléments comparés ne sont plus de même nature que pour les comparaisons requête-document ou entre documents.

La similarité entre requêtes peut se définir simplement par la proportion de termes communs aux deux requêtes. Cependant, ce principe ne prend pas en compte l'ordre des termes dans les requêtes pourtant souvent non dénué d'importance. Fitzpatrick et Dent (1997) proposent une solution basée sur la position des termes dans chaque requête. Ce principe peut être intégré dans notre fonction de calcul de score en le combinant avec le principe de proportion de termes communs de la manière suivante :

$$Score(Q, Q') = \left(\sum_i wpos_{t_i,Q} \cdot wpos_{t_i,Q'} \right) \cdot \varphi^{\frac{N_{Q,Q'}}{\min\left(N_Q, N_{Q'}\right)}} \tag{3.3}$$

où Q et Q' sont deux requêtes,
$wpos_{t_i,Q}$ est le poids associé à la position du terme t_i dans la requête Q,
$wpos_{t_i,Q'}$ est le poids associé à la position du terme t_i dans la requête Q',
$N_{Q,Q'}$ est le nombre de termes communs aux requêtes Q et Q',
N_Q est le nombre de termes distincts dans la requête Q,
$N_{Q'}$ est le nombre de termes distincts dans la requête Q',
φ est un réel positif.

L'utilisation d'une fonction de pondération associée à la position des termes permet de prendre en compte la position de manière linéaire ou non, voire de ne pas tenir compte de la position (poids égaux à 1).

Cette mesure de similarité entre requêtes peut être dérivée en une mesure de similarité entre listes de résultats. En effet, il est possible de considérer les listes de résultats comme des listes d'identifiants de documents. La similarité entre listes de résultats peut se baser sur la proportion de documents communs aux deux listes. Elle peut également se baser sur la position des documents au sein des listes de résultats. Une fonction de pondération associée à la position des documents peut également être utilisée. Cette fonction peut également être définie suivant les scores de pertinence attribués aux documents lors des recherches dans le cas où cette information est conservée. En effet, les scores de pertinence attribués aux documents en plus d'être utilisés pour ordonner les documents précisent également les pertinences estimées des documents par rapport à une requête. Un document peut avoir la même position dans deux listes de résultats correspondant à deux requêtes sans avoir le même score de pertinence.

3.2.6 Construction assistée de requête basée sur les expériences de recherche passées

Pour améliorer le processus de recherche d'information, nous proposons d'aider l'utilisateur dans l'expression de son besoin. Nous définissons une approche de construction de requête assistée. L'approche vise à aider l'utilisateur à trouver les termes adéquats pour construire sa requête ; elle se base sur la capitalisation des expériences de recherches passées. En effet, l'utilisateur est généralement laissé seul lors de la définition d'une requête. Afin de l'aider à trouver les termes définissant sa requête nous proposons d'explorer les requêtes précédentes. Les requêtes passées, notamment d'autres utilisateurs, sont visualisées en mettant en exergue les relations entre termes utilisés conjointement dans des requêtes. Grâce à cette exploration, l'utilisateur peut découvrir de nouveaux termes qu'il peut ajouter à sa propre requête.

Cette proposition a été mise en œuvre dans un prototype baptisé « QueryExplorer » (Chevalier et Hubert, 2009) où les expériences de recherche passées correspondent à des requêtes extraites de fichiers log du moteur de recherche AOL (Pass *et al.*, 2006).

3.2.6.1 Principe d'aide à la construction de requête

L'approche proposée est considérée comme une aide à la construction de requête basée sur des requêtes précédentes. Elle est intégrée au début du processus habituel de RI pour l'expression de la requête Cette intégration est non intrusive et l'utilisateur peut décider ou pas d'utiliser l'aide proposée. Le scénario général que nous proposons pour aider l'utilisateur dans sa construction de requête grâce aux expériences précédentes de recherche est illustré par la figure 3.7.

Tout d'abord, l'utilisateur propose un terme qui correspond à son besoin d'information. Le système trouve alors les variantes du terme employées dans des requêtes précédentes afin d'aider l'utilisateur à choisir le terme le plus approprié. Chaque terme proposé est associé au nombre de requêtes dans lesquelles il apparaît et les requêtes les plus utilisées comme exemple. L'utilisateur peut alors garder son propre terme ou choisir l'un de ceux proposés.

Selon le terme choisi, une représentation graphique est affichée à l'utilisateur qui peut l'explorer. Pendant cette exploration, l'utilisateur peut choisir des termes qui sont affichés pour les ajouter dans sa requête ou introduire de nouveaux termes. Chaque fois, une nouvelle représentation est calculée et affichée ensuite à l'utilisateur.

Enfin, lorsque l'utilisateur a terminé l'exploration de la représentation graphique, le moteur

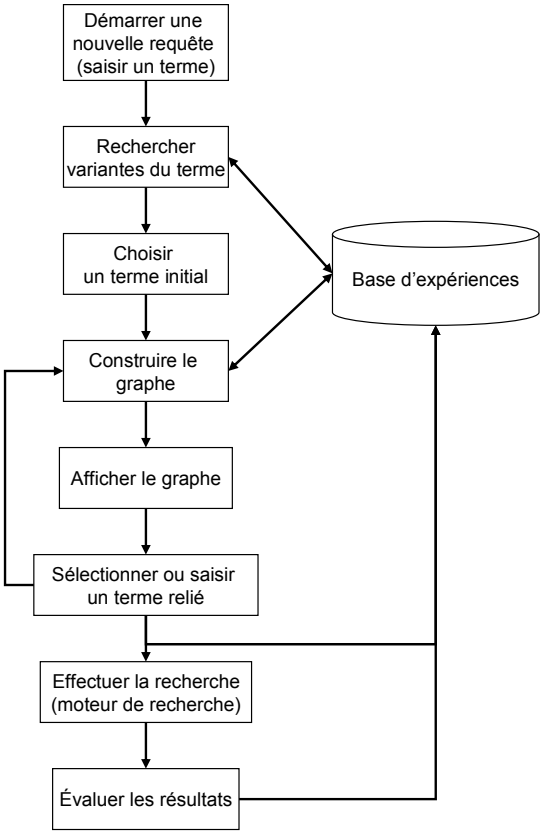


Figure 3.7 – Scénario général d'utilisation

de recherche est interrogé avec les termes choisis par l'utilisateur.

Pour explorer les termes des requêtes précédentes, nous avons identifié un ensemble minimal de fonctionnalités qui doit être associé à la représentation graphique. Pour faciliter son exploration par l'utilisateur, la représentation graphique doit :

- afficher chaque terme avec son importance. Cette importance est corrélée, dans notre approche, avec l'utilisation du terme dans les requêtes précédentes. Un terme est plus important s'il apparaît dans un plus grand nombre de requêtes ;
- afficher la relation entre les termes qui sont utilisés ensemble. Il est important de souligner les associations de termes régulièrement utilisées dans des requêtes précédentes. La visualisation peut également prendre en compte l'ordre des termes dans les requêtes précédentes ;
- comporter un ensemble d'outils de manipulation comme le filtrage (limitant les relations ayant une importance plus grande qu'un seuil par exemple) ;
- proposer différentes visualisations (points de vue) des mêmes informations, afin de tenir compte de la variabilité humaine et offrir ainsi une ou plusieurs vues compréhensibles des requêtes précédentes.

Dans notre proposition, la représentation est faite sous la forme d'un graphe. Les sommets de ce graphe correspondent aux termes des requêtes. Les arcs correspondent au lien entre deux termes, c'est-à-dire que deux termes sont reliés si et seulement s'ils ont été utilisés au moins dans une requête. Les nœuds et arcs sont pondérés. Dans le cas où peu d'informations sont disponibles, comme par exemple le cas des fichiers de log, le poids d'un nœud peut être fonction du nombre

de requêtes contenant le terme associé au nœud et le poids d'un arc peut être fonction du nombre de requêtes contenant les deux termes associés aux nœuds reliés par l'arc. Le poids d'un arc détermine son épaisseur au niveau de la visualisation. Ce graphe peut être orienté ou pas. Un graphe non orienté ne tient pas compte de la position des termes dans les requêtes passées ; la requête « t_2 t_1 » est considérée comme équivalente à la requête « t_1 t_2 » dans le calcul de l'importance du lien entre le terme t_1 et le terme t_2. Un graphe orienté tient compte de l'ordre de termes dans les requêtes passées ; l'importance du lien entre le terme t_1 et t_2 peut être différente de l'importance du lien entre le terme t_2 et le terme t_1.

3.2.6.2 Le prototype QueryExplorer

Nous avons mis en œuvre notre approche au sein d'un prototype baptisé *QueryExplorer* destiné à aider l'utilisateur à créer une requête via une exploration dans une mémoire de requêtes. Ce prototype réalisé en langage Java intègre les éléments décrits précédemment (cf. section 3.2.6.1).

Nous avons expérimenté *QueryExplorer* en utilisant un fichier log de requêtes soumises au moteur de recherche AOL (Pass *et al.*, 2006). Cette collection rassemble environ 20 millions de requêtes soumises par environ 650 000 utilisateurs sur une période de trois mois. Les données sont ordonnées suivant des identifiants d'utilisateur anonymes et de manière chronologique.

De nombreux travaux ont été effectués pour la détection de sessions de recherche dans les fichiers de logs (He et Göker, 2000; Jones et Klinkner, 2008). Nos travaux n'étant pas axés sur cette problématique, notre approche s'est appuyée sur l'étude de He et Göker (2000) qui considère comme session l'ensemble des interactions consécutives associées au même identifiant d'utilisateur et séparées par une période de temps inférieure à un seuil donné. Pour le fichier de log AOL utilisé, les doublons de requêtes dans une même session de recherche ont été supprimés pour éviter de polluer les résultats des calculs des fonctions de pondération des nœuds et des arcs. Nous conservons donc les requêtes distinctes pour chaque session de recherche. Nous avons ensuite alimenté une base de données avec les informations concernant chaque requête.

L'interface graphique utilisateur est composée de deux fenêtres principales : la première concerne la visualisation du graphe, la seconde concerne les paramètres proposés. Pour détailler ce prototype et son utilisation, la section suivante développe un scénario d'utilisation.

3.2.7 Scénario de construction d'une requête avec QueryExplorer

Pour détailler le prototype *QueryExplorer*, les figures suivantes (copies d'écran) illustrent un scénario d'utilisation. Considérons un utilisateur qui cherche des informations relatives aux triathlètes qui utilisent du matériel de marque « speedo ». L'utilisateur choisit tout d'abord le terme « speedo » pour démarrer sa recherche. À partir de ce premier terme, le système construit et affiche le graphe représentant les termes liés au terme « speedo » dans les requêtes passées comme illustré dans la figure 3.8. Le graphe représenté est orienté, c'est-à-dire que les termes « speedo, water et polo » ont été utilisés dans cet ordre dans des requêtes passées. L'épaisseur d'un arc traduit l'importance de la relation entre les deux termes ; par exemple, le couple de termes « speedo et men » a été moins souvent associés par les précédents utilisateurs que le couple « speedo et water ».

La visualisation du graphe est paramétrable suivant différentes caractéristiques modifiables dans la seconde fenêtre de paramétrage (cf. figure 3.9).

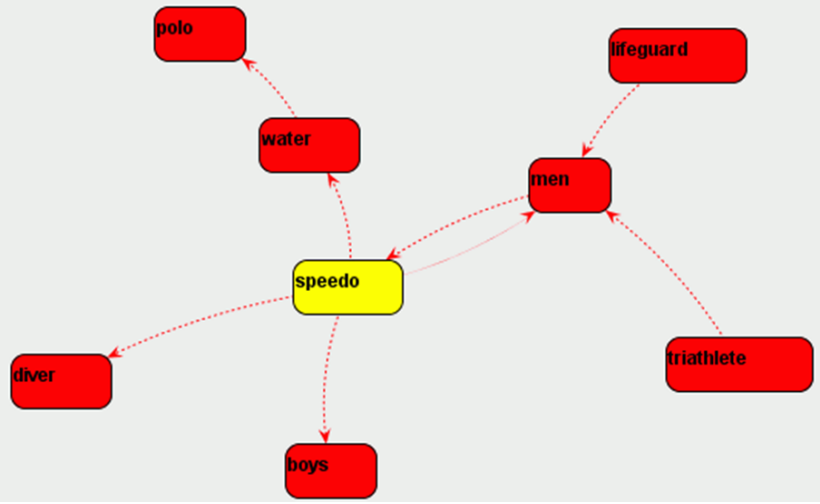


Figure 3.8 – Graphe des termes associés au terme « speedo » (Chevalier et Hubert, 2009)

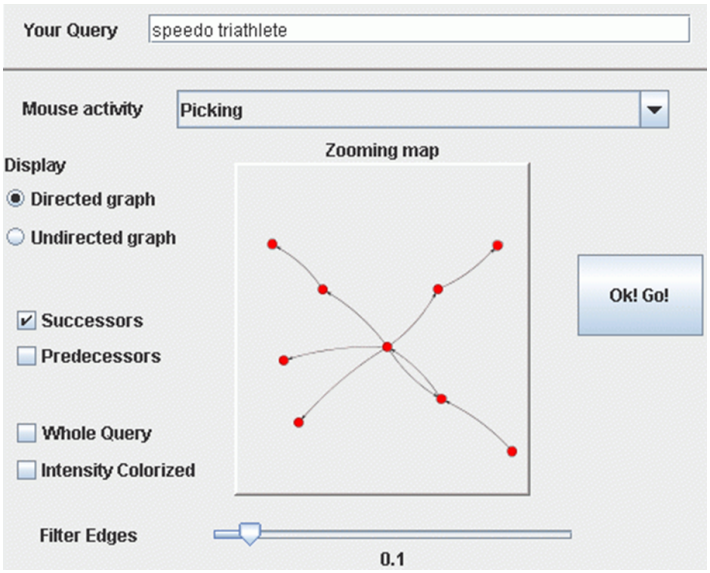


Figure 3.9 – Fenêtre de démarrage et de paramétrage de l'outil de construction de requête (Chevalier et Hubert, 2009)

Les principales caractéristiques modifiables de cette fenêtre de paramétrage sont :

- Le champ texte contenant l'expression de la requête qui peut être modifiée directement (ajout/suppression de termes) ;
- L'activité attribuée à la souris comme par exemple la sélection de termes, l'exploration du graphe ;
- Le choix de visualiser un graphe orienté ou non ;
- La possibilité de visualiser les termes successeurs et/ou prédécesseurs des termes de la requête ;
- Le choix de visualiser un graphe basé uniquement sur les requêtes passées contenant tous les termes de la requête en cours ; sinon, le graphe est basé sur les requêtes possédant au moins un terme de la requête en cours ;

– La possibilité d'ajouter au graphe l'intensité de l'importance relative des nœuds sous forme de dégradé de couleur ;
– La possibilité de filtrer les arcs. Suivant la valeur fixée, le système affiche uniquement les arcs dont le poids est supérieur au seuil fixé. De plus, seuls des nœuds reliés par des liens affichés composent le graphe.

Par exemple, si l'utilisateur décide de limiter le graphe présenté précédemment (figure 3.8) uniquement aux successeurs, les termes « lifeguard » et « triathlete » disparaissent puisqu'ils ne sont pas utilisés comme successeurs dans des requêtes passées, comme illustré dans la figure 3.10.

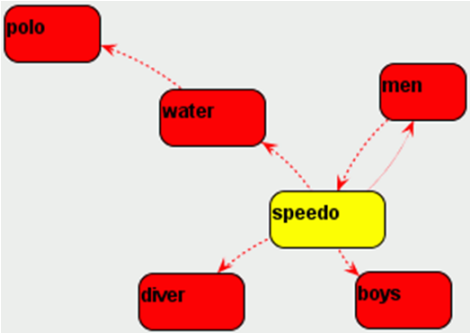


Figure 3.10 – Graphe des termes liés en tant que successeur au terme « speedo » (Chevalier et Hubert, 2009)

Si l'utilisateur ajoute un second terme à sa requête, par exemple « triathlete », de nouvelles informations peuvent apparaître comme illustré dans la figure 3.11. Certains arcs sont colorés en rouge et d'autres en noir. Les arcs rouges indiquent les termes des requêtes passées dans lesquelles apparait au moins un des termes de la requête en cours ; les arcs noirs relient d'autres termes utilisés avec les termes directement liés avec la requête en cours (mais qui n'apparaissent pas directement avec les termes de la requête en cours). Ces nœuds sont importants car ils montrent à l'utilisateur des termes en liaison avec la requête en cours bien qu'ils n'aient pas été utilisés ensemble avec des termes de cette requête auparavant. Par exemple, le terme « sc » a été utilisé avec le terme « beach » mais pas avec le terme « speedo » alors que le terme « beach » a lui été utilisé avec le terme « speedo ».

Parallèlement, l'utilisateur peut visualiser des informations associées à un nœud du graphe comme par exemple une liste des requêtes fréquentes utilisant le terme associé au nœud (figure 3.12). C'est également un moyen d'aider l'utilisateur à exprimer son besoin.

Par défaut, l'intensité des nœuds en fonction du poids associé au terme n'est pas affichée : tous les nœuds sont affichés de manière identique (exemple, figure 3.12). L'utilisateur peut décider de visualiser l'intensité de chaque nœud indiquant par exemple si le terme associé à chaque nœud est très utilisé ou non. Dans la figure 3.13, plus la teinte du nœud est foncée plus rare a été son utilisation dans des requêtes passées (par exemple le terme « sc » a moins été utilisé auparavant que le terme « beach », lui-même moins utilisé que le terme « men ».

Enfin, si l'utilisateur définit sa requête suivant des termes jamais utilisés ensemble auparavant, le graphe affiché est non connexe comme illustré dans la figure 3.14. Le graphe est alors composé de plusieurs sous-graphes correspondant à des parties de requêtes apparaissant dans des requêtes passées.

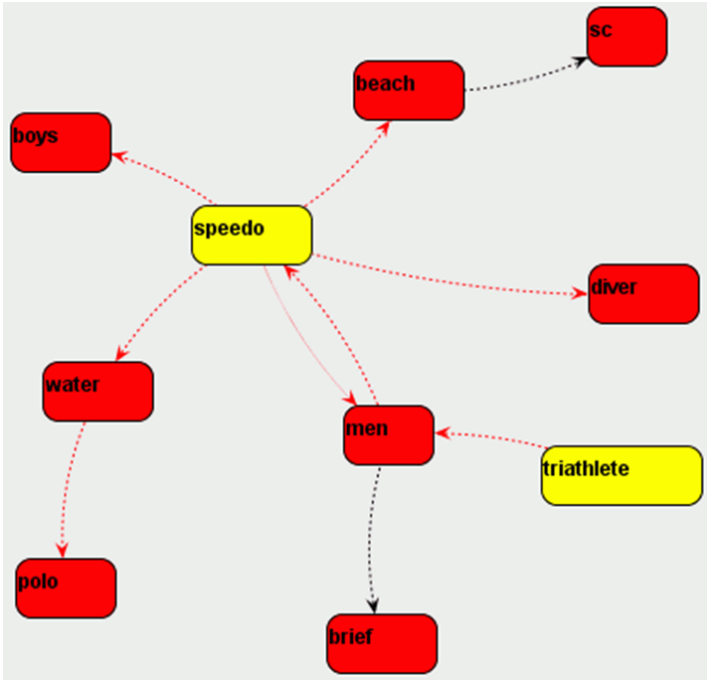


Figure 3.11 – Graphe des termes liés aux termes « speedo » et « triathlete » (Chevalier et Hubert, 2009)

Figure 3.12 – Info-bulle listant des requêtes contenant le terme

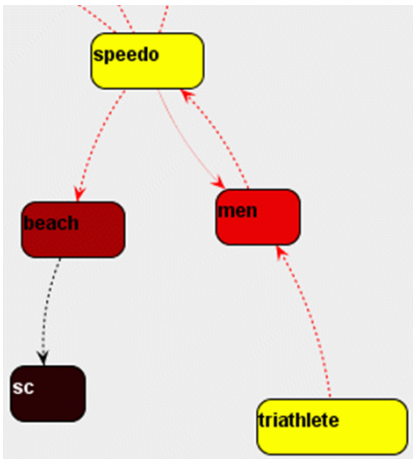


Figure 3.13 – Graphe avec intensité restituant la fréquence de cooccurrence des termes (Chevalier et Hubert, 2009)

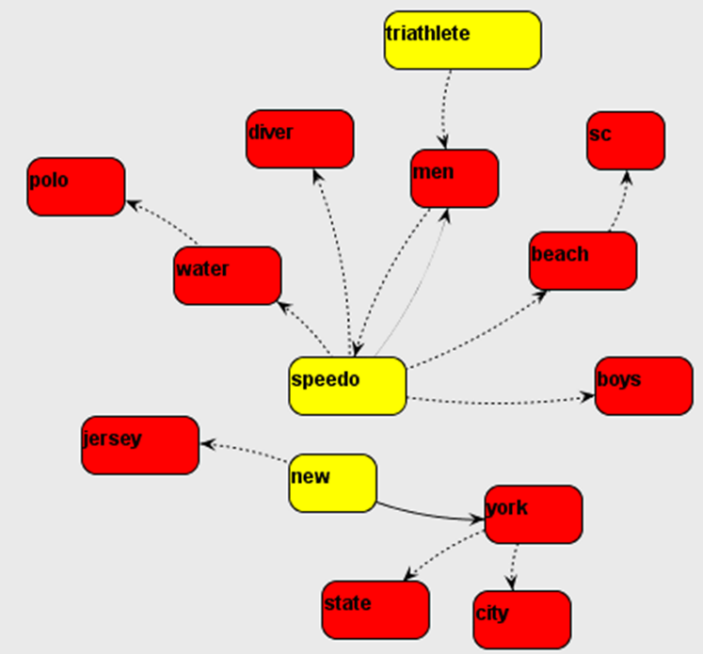


Figure 3.14 – Graphe de requête définie par des termes n'apparaissant jamais ensemble dans des requêtes passées (Chevalier et Hubert, 2009)

4 Bilan et perspectives

Ce chapitre a décrit nos contributions concernant la prise en compte de l'utilisateur dans le processus de la recherche d'information. L'objectif a été d'améliorer le processus de RI en exploitant des informations relatives à l'utilisateur et à son utilisation du SRI.

4.1 Contributions

Dans ce cadre, nos contributions ont suivi deux orientations :

– nous nous sommes intéressés, dans une première orientation, à la gestion de profils d'interrogation construits à partir des différentes sessions de recherche effectuées par un utilisateur. D'une part, nous avons proposé un modèle de gestion de profils d'interrogation reposant sur les notions de profil à court terme et de profil à long terme. D'autre part, nous avons mené des travaux sur des méthodes d'évolution du contenu des profils à partir des résultats de recherche obtenus. En effet, une problématique est de maintenir un profil qui corresponde au plus près d'un centre d'intérêt de l'utilisateur.
– nous avons proposé, dans une seconde orientation, une approche basée sur une gestion de versions pour conserver à la fois les recherches effectuées par l'utilisateur et par la même ses centres d'intérêt, ainsi que sa démarche d'interrogation. L'intérêt est de conserver une grande finesse dans les informations relatives à l'utilisateur. Ce principe offre des possibilités d'exploitation plus grandes en termes d'aide à l'utilisateur dans l'expression de ses besoins et dans l'accès aux documents correspondants.
  Dans cette orientation, nous avons également proposé une approche qui offre à un utilisateur la possibilité d'exprimer son besoin d'information en cheminant dans un graphe

construit à partir des requêtes exprimées précédemment par les autres utilisateurs. Un prototype met en œuvre cette proposition.

Ces travaux ont été menés, pour partie, en lien avec le projet européen IRAIA (cf. Annexe, section 1).

4.2 Encadrement et diffusion scientifique

Les travaux menés dans cet axe ont donné lieu à différentes publications :

Thème	Publications
Gestion de profils d'interrogation	Revue nationale : – ISDM (Benammar *et al.*, 2003) Conférences internationales : – ECIR'02 (Benammar *et al.*, 2002b) Conférence nationale : – VSST'01 (Benammar *et al.*, 2001)
Réutilisation d'expériences de recherche	Conférences internationales : – HSI'09 (Chevalier et Hubert, 2009) – ICSOFT'07 (Hubert et Mothe, 2007b)
Recherche de la nouveauté	Conférence internationale : – VSST'04 (Dkaki *et al.*, 2004)

Ces travaux ont permis le déroulement de la thèse de Anis Benammar (2003) que j'ai co-encadrée avec le professeur Josiane Mothe. Ils ont également permis de proposer des cadres d'étude de stages de Master Recherche pour deux étudiants, Mohamad El Makssoud (2010) et Randa Al Sabbagh (2005), dont j'ai assuré l'encadrement.

Ces travaux ont également conduit à des collaborations avec différents partenaires dans le cadre du projet européen IRAIA (cf. Annexe, section 1).

4.3 Perspectives

Les perspectives que j'envisage à ces travaux suivent deux orientations. Nous avons fait des propositions pour modéliser et exploiter les expériences de recherche basées sur les requêtes et les résultats de recherche associés. Outre la poursuite de ces travaux au travers du développement d'une plateforme opérationnelle, la mise en place d'une évaluation des propositions est nécessaire. Au-delà de ces perspectives directes, une orientation de nos travaux sera tournée vers l'intégration d'autres éléments de contexte comme le domaine en lien avec nos contributions sur cet aspect (cf. chapitre 1) ou la structure des documents (cf. chapitre 2). De telles informations permettront d'affiner le rapprochement entre expériences de recherche et les recommandations faites aux utilisateurs.

Une autre orientation concerne le RI sociale. Il s'agit de rapprocher les utilisateurs à partir de leurs expériences de recherche. Les utilisateurs qui partagent un certain nombre d'expériences de

recherche peuvent être connectés. Ces personnes forment donc un réseau. Il est ensuite possible d'étendre le réseau d'un utilisateur par les connexions des utilisateurs de son réseau. Une autre perspective est de rechercher les utilisateurs intéressés par un sujet c'est-à-dire qui ont effectué des expériences de recherche autour du sujet, par exemple pour trouver des collaborateurs ou des personnes ressources.

4 Adapter le système

1 Introduction

Les chapitres précédents introduisent différentes problématiques relatives à la prise en compte d'éléments du contexte dans le processus de recherche d'information, à savoir la prise en compte du domaine, de la structure de l'information. Ces approches proposent d'intégrer de nouvelles données liées au contexte de la recherche telles que des profils utilisateurs ou des ontologies de domaine dans les traitements réalisés dans le processus de RI. Ce type d'approche étend les traitements à de nouvelles données relatives au mais ces traitements restent identiques à chaque recherche. Ce chapitre traite d'une autre orientation qui consiste à chercher à adapter le comportement d'un système en modifiant les traitements réalisés en fonction de différents paramètres liés au contexte.

La structuration de ce chapitre est la suivante. La section 2 présente les propositions de la littérature autour de problématiques liées à l'adaptation de système en RI et introduit nos propositions. Les sections 3, 4, 5 et 6 détaillent nos propositions pour adapter le système lors de la recherche en fonction d'élément du contexte. La section 7 dresse le bilan de nos contributions et présent les perspectives de travaux envisagées.

2 Problématique et travaux de la littérature

Les travaux de la littérature qui visent à modifier le comportement d'un système en fonction de différents paramètres liés au contexte cherchent à répondre à différentes problématiques telles que :

- étudier et comprendre la variabilité des performances des SRI dans le traitement des différentes requêtes.
- agir sur la définition des traitements en réalisant une phase d'apprentissage pour trouver le meilleur comportement du SRI ou en combinant des SRI ayant des comportements différents,
- agir sur la restitution des résultats à l'utilisateur en proposant une visualisation appropriée du résultat fourni par le SRI. Pour cela, de nombreuses propositions d'interfaces existent et des propositions sont faites pour évaluer de leur utilisabilité.

Les sections suivantes présentent les travaux de la littérature liés à ces trois problématiques.

2.1 Variabilité des performances des SRI dans le traitement des requêtes

L'évaluation des SRI est une préoccupation de la communauté scientifique depuis les années 60 (Cleverdon, 1967). Cette préoccupation s'est poursuivie et pérennisée au travers de programmes d'évaluation annuels tels que TREC [1], CLEF [2] et INEX [3]. De manière générale, un SRI est évalué sur la base d'une collection de test comprenant un ensemble de requêtes, un ensemble de documents et pour chaque requête les documents pertinents attendus. Des mesures de similarités notamment basées sur les notions de rappel et de précision permettent de comparer les systèmes entre eux. Généralement, les systèmes sont comparés de façon globale, c'est-à-dire en calculant des valeurs moyennes obtenues pour l'ensemble des requêtes. Pourtant lorsqu'on se place au niveau de chaque requête, on observe une grande variabilité d'efficacité de traitement, d'une requête à l'autre pour un SRI donné et d'un système à l'autre.

La variabilité des performances des SRI dans le traitement des requêtes est un problème majeur. Comprendre cette variabilité pourrait permettre d'adapter le système ou de choisir le système à chaque requête. Selon Harman et Buckley (2004), comprendre cette variabilité n'est pas facile car il est difficile de séparer les défaillances liées au besoin et celles dépendantes du système. Banks *et al.* (1999) ont mené six analyses différentes sur les données TREC dans le but de dégager des relations entre les différences entre systèmes les différences dans le traitement des requêtes. Cependant aucune conclusion n'a pu être établie. Carmel *et al.* (2006) ont montré que la difficulté d'un besoin dépend de la distance entre trois composants : la description du besoin, l'ensemble des documents pertinents, et la collection de documents. Mandl et Womser-Hacker (2003) ont analysé les données de CLEF et les corrélations entre les caractéristiques des requêtes et les performances du système. Les analyses montrent une corrélation de 0,4 entre le nombre de noms propres et la précision moyenne. Dans (He et Ounis, 2003b), des caractéristiques statistiques des requêtes sont utilisées pour classifier les requêtes. Une phase d'apprentissage conduit à associer la meilleure pondération en termes de rappel/précision à chaque groupe de requêtes. Lorsqu'une nouvelle requête est soumise elle est catégorisée dans l'un des groupes existants et la pondération associée à ce groupe est appliquée pour traiter la recherche. Cette méthode a montré une amélioration des résultats pour les requêtes peu performantes de la tâche « Robustesse » de TREC (« TREC Robust Task »). Au sein de notre groupe de recherche une analyse des besoins d'information TREC selon 13 caractéristiques linguistiques a montré que la valeur moyenne de polysémie des termes des requêtes et le rappel était corrélés (Mothe et Tanguy, 2005).

Cronen-Townsend *et al.* (2002) calculent l'entropie relative entre le modèle de langage de la requête et celui de la collection pour obtenir un score de clarté (« clarity score ») qui permet de prédire la performance de traitement d'une requête. Également dans le cadre de la tâche « Robustesse » de TREC, Yom-Tov *et al.* (2004) définissent deux algorithmes de prédiction de difficulté d'une requête pour améliorer les performances du SRI. Sur la même collection, Kwok (2005) combine les *idf* (fréquence documentaire inverse) et les fréquences moyennes des termes des requêtes pour prédire les requêtes faciles et difficiles principalement pour des requêtes courtes.

Une analyse différente est menée dans (Mizzaro et Robertson, 2007) dans le but d'identifier un ensemble réduit de besoins d'information qui permettent de distinguer les systèmes efficaces

1. http://trec.nist.gov/
2. http://www.clef-campaign.org/
3. http://www.inex.otago.ac.nz/

et ceux inefficaces. Une des conclusions est que les besoins d'informations les plus faciles permettent de mieux distinguer les systèmes par rapport à leur efficacité.

En recherche d'information XML, peu de travaux se sont intéressés à l'analyse de l'efficacité des systèmes suivant des types de requêtes. Pehcevski *et al.* (2005) divisent les requêtes INEX en deux catégories « Broad » et « Narrow » suivant la spécificité (niveau article peu spécifique opposé au niveau paragraphe très spécifique) des éléments jugés majoritairement pertinents. L'étude montre que deux approches différentes sont plus efficaces pour chacune des catégories de requêtes. Sigurbjörnsson *et al.* (2005) tentent de distinguer les requêtes INEX selon le nombre d'éléments constituant les jugements de pertinence, la taille et le degré d'imbrication des éléments et ce afin d'expliquer les différences d'efficacité obtenues d'une requête à l'autre. Ces classifications possèdent néanmoins l'inconvénient de se baser sur une étude des jugements de pertinence.

2.2 Apprentissage et combinaison de SRI

Les SRI proposés par les participants aux campagnes d'évaluations telles que TREC ou INEX sont définis suivants différents paramètres permettant de modifier les résultats produits. Les variations de différents paramètres conduisent à différents résultats d'exécution d'un système soumis à évaluation dans le cadre d'une campagne. Cependant, le paramétrage est manuel et a priori, avec pour objectif d'obtenir une configuration du SRI plus efficace que les autres.

Certaines approches en recherche d'information se sont intéressées au paramétrage automatique des SRI. Par exemple, des algorithmes génétiques sont appliqués dans (Fan *et al.*, 2004a) et combinés avec de la réinjection de pertinence dans (Fan *et al.*, 2004b) pour sélectionner les pondérations de termes à utiliser et les fonctions qui les combinent. Mandl et Womser-Hacker (2005) proposent un modèle d'apprentissage qui utilise les retours de pertinence des utilisateurs pour adapter progressivement à chaque utilisateur la pondération de différents systèmes utilisés dans une fusion des résultats. He et Ounis (2003a) proposent une méthode de normalisation automatique des fréquences de termes sur lesquelles sont fondées nombre de fonctions d'appariement, dans l'esprit des normalisations proposées précédemment par (Robertson *et al.*, 1995) et (Singhal *et al.*, 1996). En recherche d'information XML peu d'approches traitent de ce point. Liu *et al.* (2004) présentent une approche permettant une indexation et un ordonnancement des résultats configurables. Cette approche a produit des résultats significatifs sur les données INEX 2003.

Cette orientation a été proposée dans le cadre de la classification de texte. Dans ce cadre, un groupe important d'approches consiste à utiliser une phase d'apprentissage pour déterminer les descripteurs qui seront utilisées pour la classification (Dash et Liu, 1997; Liu *et al.*, 2002; Cohen *et al.*, 2004; Dasgupta *et al.*, 2007; Li *et al.*, 2009).

Dans le cadre de la classification, un second type d'approche a étudié les possibilités de sélection dynamique de classifieurs (Jaimes et Chang, 2000; Mihalcea, 2002; Dimililer *et al.*, 2007).

Le principe d'apprentissage en recherche d'information connaît un nouvel intérêt notamment au travers des ateliers « Learning to Rank for Information Retrieval » adossés en 2007 et 2008 à la conférence SIGIR. L'apprentissage sur un jeu d'essai (c'est-à-dire un ensemble de requêtes leurs résultats « idéalement » ordonnés) vise à déterminer la fonction d'ordonnancement optimale qui sera ensuite appliquée pour traiter de nouvelles requêtes (Cao *et al.*, 2006; Elsas *et al.*, 2008; Taylor *et al.*, 2008; Xia *et al.*, 2008).

Enfin des travaux se sont intéressés à la combinaison de systèmes (Fox et Shaw, 1994; Lee, 1997; Montague et Aslam, 2002; Wu et McClean, 2006; Wu, 2009). Fox et Shaw (1994) ont montré que la combinaison de résultats issus de différentes stratégies de recherche améliore les performances par rapport à une stratégie appliquée seule. L'une des stratégies les plus performantes baptisée ComSUM consiste à additionner les scores de similarité des documents dans les différents ensembles fusionnés. Lee (1997) précise que l'efficacité de la fonction de combinaison CombSUM dépend du degré de recoupement des ensembles de documents pertinents et non pertinents. Il propose également de normaliser les scores de similarité des listes de résultats avant leur combinaison. Beitzel *et al.* (2003) montrent que l'amélioration est plutôt liée au nombre de documents pertinents apparaissant dans l'un des ensembles. Wu et McClean (2006) se basent sur l'analyse de corrélations entre les listes de résultats des systèmes à fusionner pour pondérer les contributions de chaque système suivant la requête. Les expérimentations menées montrent cependant seulement une légère amélioration. Des approches fondées sur un apprentissage, à l'instar de celles citées précédemment, se sont également intéressées à la combinaison de systèmes. La phase d'apprentissage sur un jeu d'essai est utilisée pour déterminer la manière optimale de combiner les résultats de différents systèmes (Joachims, 2002; Freund *et al.*, 2003; Vu et Gallinari, 2008).

2.3 Choix de visualisation des résultats de recherche

Dans le cadre de la recherche d'information une attention particulière est portée sur la dernière étape du processus qui consiste à visualiser les résultats de recherche d'une manière spécifique au travers d'une interface visuelle de recherche d'information (VIRI – « Visual Information Retrieval Interface » (Korfhage *et al.*, 1995)). Ce dernier point est très important dans le processus de RI puisqu'il est censé faciliter le traitement des résultats de recherche par l'utilisateur (trouver les documents pertinents pour lui). À l'instar des travaux qui visent à sélectionner le meilleur SRI ou à trouver la meilleure combinaison de systèmes (cf. section 2.2) le choix de la VIRI, destinée à faciliter l'accès à l'information pour l'utilisateur, est également une problématique à résoudre. Le nombre de propositions de VIRI (Chen, 2006; Hearst, 2009) confirme l'importance de ce point pour lequel il est possible de distinguer plusieurs courants.

Les VIRI traditionnelles, et probablement les plus utilisées, consistent en un affichage linéaire des résultats sous forme de liste ordonnée selon une pertinence système (par exemple, Google). Une évolution de la liste de résultats est l'affichage de groupe de documents à l'image de Grouper (Zamir et Etzioni, 1999) ou du moteur de recherche Clusty [4]. Ce dernier utilise les 100 premiers résultats pour créer les classes. La classification peut aussi être hiérarchique comme dans l'interface 2D de Grokker [5]. La classification est également utilisée dans d'autres types de VIRI, notamment les interfaces de type cartographique basées sur des cartes auto-organisatrices telles que le projet Websom (Lagus *et al.*, 1996). La visualisation cartographique peut également être basée sur les liens qui existent entre les différents résultats. C'est le cas du méta-moteur KartOO [6] dans lequel l'utilisateur visualise les liens entre les résultats, par exemple les mots-clés partagés par les résultats. Cependant, KartOO présente une succession de cartes, faisant persister l'aspect linéaire de la

4. http://www.clusty.com
5. http://www.grokker.com
6. http://www.kartoo.com

liste de résultats.

D'autres interfaces vont plus loin dans la restitution des résultats, du point de vue graphique ou métaphorique, en utilisant la spatialisation 3D des résultats de recherche. Deux principales approches se distinguent : la première approche est orientée « 3D simple » où la troisième dimension a pour objectif d'augmenter l'espace d'affichage, et la seconde approche qui peut être décrite comme un espace virtuel 3D où l'utilisateur se déplace entre les résultats.

Dans la première approche, Cat-a-Cone (Hearst et Karadi, 1997) utilise la visualisation 3D d'arbres coniques (« cone tree ») (Robertson *et al.*, 1991) pour afficher simultanément les résultats de recherche et les catégories prédéfinies correspondantes. D'autres propositions distinguent les documents selon l'importance de différents critères tels que le poids des termes de la requête comme dans les propositions de Rohrer *et al.* (1999) et Houston et Jacobson (2000) ainsi que les systèmes NIRVE (Cugini *et al.*, 2000) et Easy-DoR (Chevalier, 2002).

La seconde approche est fondée sur des métaphores 3D cartographiques à l'image de l'interface ViOS [7]. Le prototype SmartWeb (Bonnel *et al.*, 2006) utilise également cette notion pour représenter les résultats de recherche dans une ville virtuelle en 3D. Cependant, contrairement à ViOS, SmartWeb organise les résultats dans un espace 3D suivant une carte auto-organisatrice qui regroupe et place les résultats suivant la distribution des mots.

Plusieurs propositions ont été faites pour l'évaluation d'une interface. Ces approches visent essentiellement à mesurer l'utilisabilité d'une interface. Deux grandes tendances peuvent être distinguées : les méthodes « analytiques » et les méthodes « empiriques ».

Les méthodes analytiques utilisent des simulations d'activités utilisateur sans implication réelle d'utilisateur ; ces simulations sont réalisées par des experts. Les méthodes analytiques regroupent par exemple :

- l'évaluation heuristique proposée par Nielsen et Molich (1990). Cette évaluation se base sur une analyse faite par des experts qui utilisent des approches heuristiques pour identifier les problèmes ergonomiques et leur attribuer un taux de sévérité. Pour réaliser cette évaluation, Nielsen et Molich (1990) proposent neuf approches heuristiques, comme par exemple, l'utilisation d'un langage simple et naturel ou la minimisation de la charge cognitive de l'utilisateur. L'efficacité de l'évaluation heuristique dépend du nombre d'experts et de l'interface elle-même. En effet, certaines interfaces sont plus simples à évaluer avec les principes heuristiques que d'autres. De plus, le coût de ce type d'évaluation est élevé, notamment en raison du nombre d'experts nécessaires ;
- la simulation mentale (« cognitive walkthrough ») (Wharton *et al.*, 1994) qui consiste en une simulation de l'utilisation d'une interface par un utilisateur caractérisé par un profil particulier. Des experts réalisent cette simulation. Cette évaluation comporte trois étapes : la définition des données nécessaires (liste des actions, description de l'interface...), l'exécution des actions, et l'explication des résultats obtenus. Cette évaluation est efficace car elle est basée sur des actions et permet d'identifier un grand nombre de problèmes. Une limite de cette méthode réside dans le fait qu'un expert doit simuler le comportement d'un utilisateur correspondant à un profil particulier ;
- la revue des directives (« guidelines review ») ; les directives définissent les propriétés d'une

7. `http://en.wikipedia.org/wiki/ViOS`

interface (comment doit être une interface pour être performante). En vérifiant si une interface respecte les directives, des experts peuvent identifier ses lacunes. Les directives sont définies par des experts en conception d'interface. Des exemples de directives peuvent être trouvées dans (Smith et Mosier, 1986; Brown, 1988) ;
– l'inspection formelle de l'utilisabilité (Dumas et Redish, 1999) qui prône que pour parvenir à une bonne interface, des experts doivent rencontrer des utilisateurs représentatifs pour discuter des problèmes rencontrés et évolutions à apporter.

Les méthodes empiriques sont basées sur l'observation des comportements d'utilisateurs réels lorsqu'ils utilisent une interface donnée. Des exemples de méthodes empiriques sont :

– l'évaluation à base d'entretiens individuels qui visent à souligner les problèmes sévères identifiées par chaque utilisateur durant sa réelle interaction avec l'interface. Kuhn (2004) souligne également qu'il est important de faire une réunion avec tous les utilisateurs pour homogénéiser la façon dont les utilisateurs présentent les problèmes ;
– le test d'acceptabilité proposé par Shneiderman et Plaisant (2004) suivant cinq critères : durée d'apprentissage, vitesse d'exécution des actions, taux d'erreur, durée de rémanence (combien de temps l'utilisateur mémorise comment utiliser l'interface), et la satisfaction subjective ;
– l'évaluation basée sur l'observation de l'utilisateur durant son interaction avec l'interface. Cette observation peut être réalisée avec caméra et/ou microphone. L'observation audio (Nielsen et Mack, 1994) consiste à enregistrer l'utilisateur qui explique sa manipulation dans l'interface. Des experts traitent ensuite les observations pour identifier les problèmes et les évolutions possibles ;
– l'évaluation par questionnaires consiste, pour chaque utilisateur, à remplir un questionnaire. Celui proposé par Shneiderman (1997) baptisé QUIS (« Questionnaire for User Interaction Satisfaction ») recense tous les problèmes identifiables durant l'utilisation d'une interface.

Ces approches d'évaluation ont également été adaptées pour mesurer l'utilisabilité d'interfaces appliquées au processus de RI (Chevalier, 2002; Bonnel *et al.*, 2006). Plusieurs niveaux d'évaluation peuvent être considérés (Thomas et Cook, 2005). Chaque VIRI est généralement évaluée de manière spécifique, ce qui rend particulièrement difficile les comparaisons d'évaluations. Des études concernant la définition d'un cadre d'évaluation ont vu le jour (Fekete et Plaisant, 2004; Plaisant *et al.*, 2008) mais une des conclusions est que la comparaison d'interfaces reste difficile. En effet, l'utilisabilité est liée à chaque interface sans réelle possibilité de comparaison.

2.4 Orientations de nos contributions

La variabilité des SRI d'une recherche à l'autre reste une problématique assez peu abordée dans la littérature et peu de conclusions ont été établies (Banks *et al.*, 1999; Harman et Buckley, 2004). Identifier les critères qui ont une influence sur l'efficacité des SRI reste une problématique. De plus, comme le montrent des campagnes d'évaluation comme TREC ou INEX, de nombreux systèmes sont paramétrables pour obtenir des comportements différents et des résultats différents. Pourtant, rares sont les approches qui définissent comment agir sur les paramètres de ces systèmes pour obtenir les configurations les plus efficaces suivant des types de requêtes, des types de tâches de recherche ou encore des préférences utilisateurs. Les approches d'apprentissage au-

tomatique de fonction de score (Cao *et al.*, 2006; Elsas *et al.*, 2008; Taylor *et al.*, 2008) s'intéressent cependant depuis peu à cette problématique.

Dans le cadre qui consiste à considérer un processus de RI adapté en fonction de différents éléments liés au contexte, l'interface de restitution du résultat est un composant essentiel. Bien que des propositions pour l'évaluation des interfaces existent d'un point de vue utilisabilité (Julien *et al.*, 2008; Plaisant *et al.*, 2008), l'identification d'une interface restitution adaptée à chaque contexte reste une problématique.

C'est vers ces aspects que nous avons orientés d'une part nos travaux vers :

- la définition d'une approche d'adaptation de la fonction d'appariement en fonction de tâches de recherche et de préférences utilisateur ;
- la définition d'une approche de sélection SRI suivant les retours de pertinence sur les premiers documents classés par différents systèmes. Le principe est de se baser sur les jugements de pertinence de l'utilisateur sur un ensemble réduit de documents issus d'une recherche initiale afin de sélectionner le système dont les résultats seront restitués à l'utilisateur.

Dans le cadre de l'adaptation du système au contexte, nous avons orientés d'autre part nos travaux vers l'étude de l'influence de différentes caractéristiques des requêtes sur la variabilité des SRI en RI XML. Dans ce cadre, nous nous intéressons enfin à la restitution des résultats de recherche, aspect fondamental dans le processus de RI si l'on en juge par le nombre d'interfaces proposées. Notre objectif est de définir un modèle d'évaluation de l'adéquation d'une interface de restitution à un contexte de recherche et une plateforme d'évaluation. Ces travaux constituent une étape nécessaire pour la définition d'un système qui adapte la visualisation du résultat en fonction de la recherche.

3 Configurer le SRI suivant le scénario de recherche

Dans le cadre des campagnes d'évaluation telles que TREC ou INEX, différents résultats d'exécution (« runs ») d'un SRI sont soumis pour le même jeu de requêtes. Ces résultats sont obtenus en faisant varier différents paramètres intervenant par exemple dans l'indexation des documents ou encore dans l'appariement entre requêtes et documents. Cependant, le paramétrage est manuel et arbitraire dans l'objectif d'obtenir des résultats différents et au moins l'un d'entre eux qui obtienne une bonne évaluation. Il n'y a pas d'analyse de l'impact de chaque paramètre modifié sur les résultats d'évaluation obtenus. De plus, de nombreuses mesures sont proposées pour évaluer l'efficacité des systèmes. Malgré une certaine redondance entre certaines mesures (Baccini *et al.*, 2010), un ensemble d'entre elles indique différents types de performances comme par exemple la capacité à restituer globalement des documents pertinents (« précision moyenne ») ou encore la capacité à restituer des documents pertinents en tête de résultat (« haute précision »). Pourtant, aucune étude n'est réalisée pour connaître la manière de faire varier les différents paramètres pour améliorer les évaluations suivant une mesure donnée. Ceci est dû au fait que dans le cadre des campagnes d'évaluation, une mesure d'évaluation concentre toute l'attention, généralement la MAP (« Mean Average Precision »). L'objectif des participants est alors focalisé sur l'obtention des meilleurs résultats pour cette mesure.

Dans l'optique d'un système qui s'adapte à chaque recherche, il semble pourtant important de savoir comment agir sur la définition du système pour obtenir la meilleure efficacité suivant un scénario donné. C'est pourquoi nous avons mené ce type d'étude sur notre SRI dans le cadre des campagnes d'évaluation INEX. Ce cadre est intéressant pour notre étude car il procure différents scénarios de recherche au travers de la définition de différentes quantifications qui simulent différentes préférences utilisateurs (cf. chapitre 2, section 3.6.1.1). Nous avons donc étudié l'influence de certains paramètres sur les résultats d'évaluations pour différents scénarios afin de déduire comment faire varier ces paramètres pour améliorer l'efficacité de notre SRI pour un scénario donné.

3.1 Possibilités d'adaptation du SRI

La base de cette étude est le modèle de RI XML détaillé au chapitre 2, dont les principales définitions sont rappelées ci-dessous. Le principe général de calcul de score d'un élément B par rapport à un élément A est défini selon la fonction suivante (cf. chapitre 2, section 3.1, équation 2.1) :

$$Score(A,B) = \left(\sum_{c} imp_{BI}(c,A) \cdot imp_{UI}(c,B) \cdot corr_{struc}(c,A,B) \right) \cdot rec_{cont}(A,B) \cdot rec_{struc}(A,B)$$

où A et B sont des éléments (éventuellement représentés différemment)
c est un concept représentant A ou B.

- $imp_{BI}(c,A)$ est le facteur qui évalue l'importance du concept c_i pour l'élément A.
- $imp_{UI}(c,B)$ est le facteur qui évalue l'importance du concept c_i pour l'élément B.
- $corr_{struc}(c,A,B)$ est le facteur évalue la correspondance entre l'information de structure associée au concept c_i dans l'élément A et celle associée à cet élément dans l'élément B.
- $rec_{cont}(A,B)$ est le facteur qui évalue l'importance du recouvrement de contenu entre l'élément A et l'élément B.
- $rec_{struc}(A,B)$ est le facteur qui évalue l'importance du recouvrement de structure entre l'élément A et l'élément B.

Une fonction de score appliquée à la RI XML est alors la suivante (cf. chapitre 2, sections 3.3 et 3.4) :

$$Score_a(Q,E) = cg_{Q,E} \cdot c_{Q,E} \cdot \left(Score(Q,E) + \sum_{l} \alpha^{\frac{d(E,E_l)}{d(E_r,E_l)}} \cdot Score(Q,E_l) \right) \tag{4.1}$$

avec

$$Score(Q,E) = \left(\sum_{i} ct_{i,Q,E} \cdot pf_{i,Q} \cdot \frac{f_{i,Q}}{ef_i} \cdot tf_{i,E} \right) \cdot c_{Q,E} \cdot \varphi^{\frac{N_{Q,E}}{\min(N_Q,N_E)}} \tag{4.2}$$

$$Score(Q,E_l) = \left(\sum_{j} ct_{j,Q,E} \cdot pf_{j,Q} \cdot \frac{f_{j,Q}}{ef_j} \cdot tf_{j,E_l} \right) \cdot c_{Q,E} \cdot \varphi^{\frac{N_{Q,E_l}}{\min(N_Q,N_{E_l})}} \tag{4.3}$$

où $ct_{i,Q,E} = 1 - \beta(1-a)$ tel que
Si l'élément E ne correspond pas à la contrainte sur le terme t_i dans la requête Q
alors $a = 0$ Sinon $a = 1$
β est un réel positif tel que $0,0 \leqslant \beta \leqslant 1,0$

$cg_{Q,E} = 1 - \gamma(1-a)$ tel que
Si l'élément E ne correspond pas à la contrainte de granularité de la requête Q
alors $a = 0$ Sinon $a = 1$
γ est un réel positif tel que $0,0 \leqslant \gamma \leqslant 1,0$

$c_{Q,E}$ tel que
Si $\frac{N_{Q,E}}{\min(N_Q,N_E)} \geqslant CT$ Alors $c_{Q,E} = 1,0$ Sinon $c_{Q,E} = 0,0$
où CT est un réel positif représentant le seuil tel que $0,0 \leqslant CT \leqslant 1$,
$N_{Q,E}$ est le nombre de termes communs à la requête Q et à l'élément XML E
N_Q est le nombre de termes distincts dans la requête Q
N_E est le nombre de termes distincts dans l'élément XML E

En plus d'offrir la possibilité de changer les définitions des fonctions imp_{BI}, imp_{UI}, $corr_{struc}$, rec_{cont} et rec_{struc}, l'appariement entre requête et éléments XML utilise différents paramètres φ, α, β, γ et CT. Faire varier les valeurs de ces coefficients constitue différentes possibilités d'adapter le SRI :

- la fonction qui mesure le recouvrement entre la requête et l'élément XML dépend de la valeur donnée au coefficient φ,
- l'importance de l'agrégation de score dépend de la valeur donnée au coefficient α,
- la valeur donnée au taux de couverture de la requête CT peut varier,
- les valeurs associées aux préférences sur les termes pour la fonction de pondération $pf_{i,Q}$ peuvent varier,
- la prise en compte des conditions sur la structure dépend des valeurs données aux coefficients β et γ associés aux deux types de contraintes liées à la structure XML (sur le contenu et sur le résultat).

Le SRI est défini suivant différents paramètres correspondant à différents aspects de la recherche XML. Ceci vise à permettre l'adaptation du système à différents scénarios de recherche. Les scénarios de recherche peuvent être liés à différents critères :

- la taille des éléments retournés (éléments cibles) : un utilisateur peut être intéressé en priorité par des éléments courts tandis qu'un autre utilisateur préférera en premier lieu des éléments plus longs ;
- la couverture de la requête : un utilisateur peut préférer obtenir en résultat uniquement des éléments traitant de la majeure partie des concepts exprimés dans la requête. À l'inverse, un utilisateur acceptera en résultat des éléments couvrant moins de concepts de la requête. Cependant le respect d'une couverture minimum de la requête peut être vérifié en fixant une valeur non nulle au taux de couverture (cf. chapitre 2, section 3.3.2.2) ;
- les préférences sur les concepts : un utilisateur peut être plus exigeant sur la présence de certains concepts par rapport à d'autres. De plus, il peut considérer que ces préférences doivent

être respectées avec plus ou moins de force. L'influence de chaque type de préférence sur les concepts est définie par les coefficients associés aux différentes préférences ;
- la focalisation de l'élément : un utilisateur peut être plus intéressé par des éléments se focalisant sur un concept de la requête plutôt que par des éléments traitant partiellement de tous les concepts et vice versa ;
- la vérification des indications de structure : un utilisateur peut vouloir uniquement les éléments qui répondent strictement aux contraintes de structure qu'il a indiquées ou il peut simplement préférer les éléments qui respectent ces contraintes sans considérer celles-ci comme indispensables. La force de vérification des contraintes de structure varie suivant les valeurs des coefficients qui leurs sont associés. L'importance accordée à chaque type de contrainte varie de manière indépendante. Il est possible de vérifier de manière plus ou moins stricte et différemment les contraintes sur les éléments cibles et les contraintes sur le contenu des éléments.

3.2 Expérimentations relatives à la configuration du SRI selon le scénario de recherche sur la collection IEEE - INEX 2005

L'apprentissage réalisé via la participation aux éditions 2004 et 2005 du programme d'évaluation INEX a conduit à identifier une configuration de notre SRI plus adaptée au contexte INEX notamment au regard des caractéristiques de la collection. De manière globale, cette configuration a donné de bons résultats d'évaluation. Nous avons choisi d'utiliser cette configuration comme base pour des études complémentaires visant à évaluer l'influence de certains paramètres sur le comportement de notre SRI. De plus, ces études complémentaires ont eu pour objectif d'évaluer la capacité de notre SRI à être adapté à différents types de recherches. Les expérimentations ont été réalisées avec la collection de test INEX 2005.

Les paramètres étudiés lors de ces expérimentations ont été le coefficient de recouvrement entre élément et requête et le coefficient d'agrégation de score. Ces deux paramètres ont été choisis pour leur rôle majeur dans le calcul du score d'un élément. Nous avons tenté d'évaluer l'influence de ces paramètres au regard de différentes quantifications et tâches définies dans le cadre INEX 2005 simulant différents scénarios de recherche.

Les résultats obtenus (« runs ») suivant les différentes configurations testées sont nommés de manière à identifier les valeurs choisies pour les paramètres étudiés comme suit :

- φY indique la valeur Y du coefficient φ de recouvrement entre requête et élément,
- αZ indique la valeur Z du coefficient α utilisé lors de l'agrégation de score.

Le premier volet de ces expérimentations (Tableau 4.1 et Tableau 4.2), a consisté à mesurer la possibilité de mieux répondre à un scénario de recherche en faisant varier l'agrégation des scores (coefficient α). Compte tenu de la définition du calcul de score (cf. section 3.1), une faible agrégation favorise les nœuds feuilles des documents XML ou les parents proches des nœuds feuilles. La première hypothèse était donc que, dans le contexte INEX, ceci se traduit par la promotion des éléments hautement spécifiques et tend à obtenir de meilleurs résultats pour la quantification *strict* (cf. chapitre 2, section 3.6.1.1).

De plus, accroître l'agrégation tend à inclure des nœuds plus hauts dans la structure hiérarchique des documents XML. La seconde hypothèse était donc que dans le contexte INEX ceci

correspond à inclure dans les résultats des nœuds moins spécifiques et donc d'obtenir de moins bonnes évaluations pour la quantification *strict* mais de meilleures évaluations pour la quantification *generalised*(cf. chapitre 2, section 3.6.1.1). Cependant, il est aussi prévisible qu'une trop forte agrégation conduise à une détérioration des résultats pour les deux quantifications puisque dans ce cas des éléments hautement spécifiques pourraient disparaître au profit d'éléments moins spécifiques. Les expérimentations ont montré le seuil à partir duquel ce phénomène se produit.

Mesure : ep/gr (MAep), Tâche : Thorough		
	Quantification	
Run	strict	generalised
$\alpha 0.1\varphi 1$	**0,038**	0,037
$\alpha 0.3\varphi 1$	0,031	0,052
$\alpha 0.5\varphi 1$	0,017	**0,062**
$\alpha 0.7\varphi 1$	0,011	0,056
$\alpha 0.9\varphi 1$	0,008	0,047
	Différences significatives	

Tableau 4.1 – Influence de l'agrégation de score sans facteur de recouvrement φ pour la tâche Thorough (Hubert *et al.*, 2007c)

Mesure : ep/gr (MAep), Tâche : Thorough		
	Quantification	
Run	strict	generalised
$\alpha 0.1\varphi 1000$	**0,033**	0,047
$\alpha 0.3\varphi 1000$	0,031	0,059
$\alpha 0.5\varphi 1000$	0,028	*0,066*
$\alpha 0.7\varphi 1000$	0,023	**0,067**
$\alpha 0.9\varphi 1000$	0,016	0,063
	Différences significatives	Différences significatives excepté entre $\alpha 0.5\varphi 1000$ et $\alpha 0.7\varphi 1000$

Tableau 4.2 – Influence de l'agrégation avec un fort coefficient de recouvrement φ pour la tâche Thorough (Hubert *et al.*, 2007c)

Ces expérimentations confirment qu'accroître l'agrégation de score détériore les résultats pour la quantification *strict* quel que soit le coefficient de recouvrement entre requête et élément. Pour la quantification *generalised*, accroître l'agrégation de score améliore les résultats avant de les détériorer lorsqu'elle devient trop prononcée. Les meilleurs résultats semblent être obtenus pour un coefficient d'agrégation α entre 0,5 et 0,7. L'application d'un test de significativité (t-test pairé, $\alpha = 0,05$) montre que les différences de résultats sont significatives excepté entre les résultats $\alpha 0.5\varphi 1000$ et $\alpha 0.7\varphi 1000$ pour la quantification *generalised*.

Le second volet de ces expérimentations, dont les résultats sont synthétisés dans les ta-

bleaux 4.3 et 4.4, a eu également pour but d'évaluer la possibilité de mieux répondre à un scénario donné en faisant varier le coefficient de recouvrement φ entre requête et élément. Compte tenu de la définition du calcul de score (cf. section 3.1), un coefficient de recouvrement élevé favorise les nœuds XML qui contiennent plusieurs des termes de la requête même si ceux-ci ne sont pas fortement discriminants, au détriment des éléments possédant peu de termes de la requête mais fortement discriminants. Néanmoins, les éléments contenant de nombreux termes de la requête incluant des termes fortement discriminants obtiennent dans tous les cas les scores les plus élevés. Dans le contexte INEX, augmenter le coefficient de recouvrement implique favoriser les éléments exhaustifs.

Pour la quantification *generalised*, l'hypothèse est que les résultats doivent être meilleurs en appliquant un facteur de recouvrement puisque davantage de nœuds exhaustifs seront mieux classés. Cependant, un coefficient de recouvrement trop élevé peuvent détériorer les résultats puisque trop d'éléments relativement exhaustifs mais peu spécifiques remplacent des éléments moins exhaustifs mais plus spécifiques. Pour la quantification *strict*, l'hypothèse est que le facteur de recouvrement peut compenser l'élimination de nœuds spécifiques causée par une trop forte agrégation de score. Avec une faible agrégation le facteur de recouvrement ne doit pas améliorer les résultats ayant tendance à favoriser des nœuds partiellement exhaustifs et peu spécifiques au détriment de nœuds moins exhaustifs mais plus spécifiques.

Mesure : ep/gr (MAep), Tâche : Thorough		
	Quantification	
Run	strict	generalised
α0.6φ1	0,013	0,060
α0.6φ50	0,020	**0,069**
α0.6φ500	*0,024*	*0,068*
α0.6φ1000	*0,024*	*0,068*
α0.6φ3000	**0,025**	0,067
	Différences significatives excepté entre α0.6φ500, α0.6φ1000 et α0.6φ3000	Différences significatives excepté entre α0.6φ50, α0.6φ500 et α0.6φ1000

Tableau 4.3 – Influence du facteur de recouvrement avec une agrégation de score α moyenne pour la tâche Thorough (Hubert *et al.*, 2007c)

Mesure : ep/gr (MAep), Tâche : Thorough		
	Quantification	
Run	strict	generalised
$\alpha 0.1\varphi 1$	**0,038**	0,037
$\alpha 0.1\varphi 50$	0,034	0,032
$\alpha 0.1\varphi 500$	0,033	0,036
$\alpha 0.1\varphi 1000$	0,033	0,047
$\alpha 0.1\varphi 3000$	0,032	**0,048**
	Différences non significatives	Différences significatives excepté $\alpha 0.1\varphi 50$ et $\alpha 0.1\varphi 500$

Tableau 4.4 – Influence du facteur de recouvrement avec une faible agrégation de score α pour la tâche Thorough (Hubert *et al.*, 2007c)

Les résultats confirment qu'avec une agrégation effective (Tableau 4.3) augmenter le facteur de recouvrement conduit à de meilleurs résultats d'évaluation pour la quantification *strict*. Pour la quantification *generalised* un faible facteur de recouvrement conduit aux meilleurs résultats. Augmenter ce facteur semble entraîner une faible détérioration des résultats.

Avec une faible agrégation de score (Tableau 4.4) les résultats sont meilleurs pour la quantification *strict* sans facteur de recouvrement. À l'opposé, pour la quantification *generalised* augmenter le facteur de recouvrement conduit à de meilleures évaluations.

Les tests de significativité (t-test pairé, $\alpha = 0,05$) montrent, pour la quantification *generalised*, des différences significatives entre les résultats exceptés entre les résultats $\alpha 0.1\varphi 50$ et $\alpha 0.1\varphi 500$ ainsi qu'entre les résultats $\alpha 0.6\varphi 50$,$\alpha 0.6\varphi 500$ et $\alpha 0.6\varphi 1000$. Pour la quantification *strict*, les tests de significativité montrent des différences significatives seulement pour les résultats de type $\alpha 0.6$ excepté entre les résultats $\alpha 0.6\varphi 500$,$\alpha 0.6\varphi 1000$ et $\alpha 0.6\varphi 3000$.

Au regard des expérimentations faisant varier des paramètres majeurs et des quantifications étudiées simulant différents scénarios de recherche, il est possible d'identifier comment configurer le SRI pour chaque cas. Si l'on cherche seulement des éléments XML hautement exhaustifs et totalement spécifiques (c.-à-d. scénario ⟨ tâche *Thorough*, quantification *strict* ⟩) la configuration adaptée pour le SRI est une faible agrégation de score ($\alpha = 0,1$) et pas de facteur de recouvrement ($\varphi = 1$). En revanche, si l'on cherche tous les éléments XML quel que soit leur niveau de pertinence (c.-à-d. scénario ⟨ tâche *Thorough*, quantification *generalised* ⟩) la configuration adaptée est une agrégation moyenne ($\alpha = 0,6$) et un faible facteur de recouvrement ($\varphi = 50$).

En vue de valider ces hypothèses, des expérimentations complémentaires ont été réalisées sur les requêtes de type CAS définies pour l'édition INEX 2005 suivant les critères d'évaluation de la tâche VVCAS. Cette tâche a été choisie en raison des critères d'évaluation similaires à ceux de la tâche *Thorough* pour les requêtes de type CO utilisées pour les premières expérimentations. Cette seconde série d'expérimentations a montré un comportement similaire du SRI vis-à-vis des paramètres et des scénarios étudiés. Au regard des quantifications, les configurations conduisant aux meilleurs résultats ont été les mêmes que lors des expérimentations menées sur les requêtes de type CO.

4 Choisir le SRI suivant le retour de pertinence utilisateur

Dans le domaine de l'évaluation des SRI et plus particulièrement vis-à-vis de la variabilité des performances des SRI (cf. section 2.1), l'étude de phénomènes locaux a montré que les SRI ne sélectionnent pas les mêmes documents en réponse à une requête. Les mécanismes de combinaison de résultats (Fox et Shaw, 1994) exploitent cette complémentarité dans les listes documents restitués par différents systèmes. L'hypothèse sous-jacente est qu'en combinant les réponses de différents systèmes, il est possible d'améliorer la sélection des documents et donc les performances globales. Les mécanismes retenus reposent sur une combinaison des réponses de différents systèmes pour une requête donnée. L'étude des phénomènes locaux a également montré qu'il existait une spécificité des systèmes et des requêtes. En effet, même si certains besoins sont difficiles à satisfaire pour une majorité de systèmes (Dkaki *et al.*, 2004; Harman et Buckley, 2004), certains systèmes ont des comportements spécifiques qui leur permettent de mieux répondre à certaines requêtes. C'est également ce que nous avons constaté avec notre SRI (cf. section 5).

Par exemple, le tableau 4.5 indique les performances locales (AP) sur deux besoins d'informations (« topics ») et globales (MAP) sur l'ensemble des besoins d'informations pour deux exécutions (« runs ») de systèmes différents. Ces performances sont extraites des résultats de la tâche *ad hoc* de la 7e édition de TREC (TREC7).

	Topic		
Run	352 (AP)	354 (AP)	Tous (MAP)
CLARIT98COMB	0,5068	0,1675	0,3702
T7miti1	0,3081	0,2767	0,3675

Tableau 4.5 – Performances locales et globales de deux systèmes de la tâche TREC 7 ad hoc

Au regard des performances des deux systèmes illustrées dans le tableau 4.5, s'il était possible de décider a priori d'utiliser le système CLARIT98COMB pour la requête 352 et le système T7miti1 pour la requête 354, il serait possible d'optimiser la performance globale.

Notre étude s'inscrit dans ce cadre (Hubert et Mothe, 2007a; Hubert *et al.*, 2007a). Nous nous appuyons sur le fait que les systèmes n'obtiennent pas localement les mêmes performances. Même si deux systèmes sont globalement équivalents sur un ensemble de requêtes, par effet de lissage de la moyenne, les systèmes n'obtiennent pas des performances semblables lorsque chaque requête est considérée de façon indépendante. Ainsi, notre hypothèse est qu'un système basé sur un ensemble de SRI et capable de choisir le système qui traitera au mieux la requête obtiendra des performances globales meilleures que chacun des SRI pris séparément.

Notre hypothèse, que nous testons dans ce chapitre, est que le choix du « meilleur » système peut être basé sur l'analyse des premiers documents restitués par le SRI. Ainsi, une première recherche est réalisée par chacun des SRI et, en fonction du jugement de pertinence sur les premiers documents restitués, l'un ou l'autre des systèmes sera utilisé pour traiter la requête et restituer les documents.

L'hypothèse est donc qu'il est possible de combiner judicieusement différents SRI et de sélectionner celui qui sera le plus performant pour une requête donnée. Pour valider cette hypothèse,

nous avons réalisé une expérimentation qui a consisté à utiliser différents systèmes pour chacune des requêtes et à sélectionner « manuellement » le système le plus performant pour chaque requête. Nous maximisons donc ici les résultats qu'il est possible d'obtenir par le principe de sélection de système que nous proposons. Nous avons pu montrer que dans le cas de l'utilisation des deux meilleurs systèmes ayant participé chaque année à la tâche *ad hoc* de TREC, si un système était capable de choisir toujours le « meilleur » d'entre eux, les performances en termes de MAP seraient améliorées de 10 à près de 20 %, en fonction des années par rapport au meilleur système. Lorsque les 5 meilleurs systèmes sont choisis, cette amélioration varie de 12 à 30 % (voir tableaux 4.7 et 4.8, colonne « Optimal »). Cette expérimentation justifie notre hypothèse puisqu'il est potentiellement possible d'augmenter les performances d'un système de recherche d'information en se basant sur différents SRI et en sélectionnant celui qui est le plus adapté en fonction de la requête traitée. Il est évident qu'une telle approche n'est concevable que s'il est possible de détecter a priori le système à utiliser.

La méthode que nous proposons vise donc à sélectionner le SRI utilisé pour traiter une requête sur la base de la précision à 5 documents (P@5). La P@5 est la précision calculée après 5 documents restitués par le SRI. Une série d'expérimentations appliquant cette méthode valide notre hypothèse.

4.1 Principe de sélection de SRI basée sur un retour de pertinence réduit

Nous proposons de baser la sélection du système sur des informations minimales concernant la pertinence des premiers documents retrouvés. Dans notre cas, cette information sur la pertinence des premiers documents est utilisée pour choisir le système à considérer. Nous utilisons la précision à 5 documents pour choisir parmi les systèmes celui qui traitera la requête pour l'utilisateur comme illustré dans l'algorithme 1.

pour chaque *requête Q* **faire**
  pour chaque *système S* **faire**
    Rechercher les documents pertinents
    Restituer les 5 premiers documents à l'utilisateur pour jugement de pertinence
    Calculer la précision des 5 premiers documents (P@5)
  fin
  Ordonner les systèmes par P@5 décroissante
  Choisir le premier système pour traiter la requête
  Restituer les documents retrouvés par ce système
fin

Algorithme 1: Algorithme de sélection du SRI pour chaque requête

Cette approche de sélection a été évaluée sur les collections de quatre éditions TREC *ad hoc*. Dans un premier temps, nous avons considéré pour chaque édition les deux meilleurs systèmes puis dans un second temps les cinq meilleurs systèmes pour la collection considérée. Pour une édition, les meilleurs systèmes sont ceux qui ont obtenu les meilleurs résultats globaux mesurés en termes de MAP.

4.2 Expérimentations relatives à la sélection de SRI sur des collections TREC

4.2.1 Collections et critères d'évaluation

Les évaluations que nous avons réalisées s'appuient sur les données issues des tâches *ad hoc* de quatre éditions de TREC (TREC 3, 5, 6 et 7). Pour chaque édition, ces données regroupent :

- l'ensemble des documents correspondant à la collection à interroger,
- l'ensemble des requêtes (généralement 50 pour une tâche et une année donnée),
- l'ensemble des documents que le système idéal aurait dû retrouver (jugés manuellement),
- pour chaque système participant, la liste des documents restitués pour chaque requête.

Les caractéristiques des données utilisées sont indiquées dans le tableau 4.6.

Caractéristique	TREC3	TREC5	TREC6	TREC7
Documents	≈ 1 Go	≈ 2 Go	≈ 2 Go	≈ 2 Go
Besoins d'information («Topics»)	151 à 200	251 à 300	301-350	351-400
Nombre de résultats («runs») soumis par les participants	48	77	79	103

Tableau 4.6 – Caractéristiques des données TREC ad hoc utilisées dans les expérimentations (Hubert et Mothe, 2007a)

Nous avons utilisé le programme trec_eval, utilisé dans les évaluations du programme TREC, pour calculer les valeurs des mesures globales les plus utilisées pour comparer différents systèmes de précision notamment la précision globale moyenne MAP (« Mean Average Precision »).

4.2.2 Sélection à partir des 2 meilleurs systèmes

Dans cette expérimentation, nous avons sélectionné les 2 systèmes qui ont obtenu la meilleure valeur de MAP pour une année donnée de TREC. En fonction de la valeur de la précision à 5 documents (pour chaque requête) l'un ou l'autre des systèmes est utilisé. Le tableau 4.7 présente les résultats. Dans ce tableau, la colonne « Optimal » indique la valeur de MAP si, pour chaque requête, le meilleur système avait été sélectionné. La colonne « SélectionP@5 » correspond aux résultats obtenus par notre approche de sélection à partir de la précision à 5 documents. Les résultats sont indiqués pour les 5 premières requêtes de chaque collection ainsi que sur l'ensemble des requêtes (Global). Les chiffres entre parenthèses indiquent les différences en termes de % par rapport au meilleur système sur la collection considérée.

Globalement, quelle que soit la collection, l'amélioration des performances globales en termes de MAP par la technique de sélection proposée est proche des 10 %. Si l'on considère les 5 premières requêtes de chaque collection, dans la majorité des cas, la sélection est pertinente. Les performances sont plus nuancées pour la collection TREC6. Ces résultats peuvent s'expliquer par les très bonnes performances initiales cette année là du meilleur système (0,4631 de MAP).

Il est également intéressant de noter que la différence entre les MAP des systèmes initiaux ne permet pas de prédire la pertinence de la sélection. Par exemple, pour la collection TREC5, la différence entre les résultats obtenus par les deux meilleurs systèmes ETHme1 et Uwgcx1 est de

TREC3	Inq102 (1er)	Citya1 (2e)	Optimal	SélectionP@5
Local	0,6259	0,5783	0,6259	0,6259
(5 premières	0,2699	0,5667	0,5667	0,5667
requêtes)	0,1806	0,2681	0,2681	0,2681
	0,7372	0,7354	0,7372	0,7372
	0,2504	0,0035	0,2504	0,2504
Global	0,4226	0,4012	0,4647	0,4576
			(+ 9,96 %)	(+ 8,28 %)
TREC5	ETHme1 (1er)	Uwgcx1 (2e)	Optimal	SélectionP@5
Local	0,0673	0,2215	0,2215	0,2215
(5 premières	0,0453	0,0932	0,0932	0,0453
requêtes)	0,6813	0,8600	0,8600	0,6813
	0,3262	0,2909	0,3262	0,3262
	0,1660	0,0543	0,1660	0,1660
Global	0,3165	0,3098	0,3900	0,3684
			(+ 23,22 %)	(+ 16,40 %)
TREC6	uwmt6a0 (1er)	CLAUG (2e)	Optimal	SélectionP@5
Local	0,3185	0,4753	0,4753	0,4753
(5 premières	0,7671	0,5819	0,7671	0,7671
requêtes)	0,6556	0,6779	0,6779	0,6556
	0,5000	0,2599	0,5000	0,2599
	0,0302	0,0600	0,0600	0,0600
Global	0,4631	0,3742	0,5079	0,4773
			(+ 9,67 %)	(+ 3,04 %)
TREC7	CLARIT98COMB (1er)	T7miti (2e)	Optimal	SélectionP@5
Local	0,7112	0,8366	0,8366	0,7112
(5 premières	0,5068	0,3081	0,5068	0,5068
requêtes)	0,4281	0,3388	0,4281	0,3388
	0,1675	0,2767	0,2767	0,1675
	0,4555	0,5429	0,5429	0,5429
Global	0,3702	0,3675	0,4341	0,4069
			(+ 17,26 %)	(+ 9,91 %)

Tableau 4.7 – MAP locale et globale après fusion des deux meilleurs systèmes (Hubert et Mothe, 2007a)

0,0067, pourtant, leur combinaison est très efficace (potentiellement elle est de plus de 23 % et notre méthode permet d'augmenter les performances de plus de 16 %. À l'opposé, pour la collection TREC6, la différence entre les résultats obtenus par les deux meilleurs systèmes uwmt6a0 et CLAUG est de 0,0889 mais leur combinaison ne permet qu'une amélioration de moins de 10 % potentiellement et de 3 % par notre méthode.

4.2.3 Sélection à partir des 5 meilleurs systèmes

Dans cette expérimentation, nous avons répété l'expérimentation précédente, sur la base des 5 systèmes qui ont obtenu la meilleure valeur de MAP globale pour la collection considérée. En fonction de la valeur de la précision locale à 5 documents (pour chaque requête), l'un ou l'autre des systèmes est utilisé. Le tableau 4.8 présente les résultats en utilisant les mêmes notations que dans le tableau 4.7.

Lorsque cinq systèmes sont utilisés plutôt que deux, le potentiel d'augmentation augmente quelle que soit la collection (de 10 à 14 % pour TREC3, de 17 à 30 % pour TREC7). L'amélioration des résultats par la méthode de sélection en revanche n'est pas aussi efficace que son potentiel, même si globalement, 5 systèmes permettent tout de même d'améliorer un peu plus les résultats (+ 9,40 % de la MAP par rapport au meilleur système en moyenne sur les 4 collections avec 2 systèmes et + 9,93 % avec 5 systèmes).

Ces évaluations ont été menées sur les résultats complets incluant les 5 premiers documents dont la pertinence est connue et utilisée pour la combinaison des systèmes. Il serait souhaitable de réaliser les évaluations sur la collection résiduelle c'est-à-dire en supprimant les 5 premiers documents déjà jugés.

5 Distinguer les requêtes pour choisir le SRI

Au regard des expérimentations menées dans le cadre de campagnes d'évaluations comme INEX[8] pour la recherche d'information XML ou TREC[9] pour la recherche d'information, il est clair que les systèmes ne parviennent pas à traiter avec la même efficacité toutes les requêtes proposées.

Dans le cadre INEX, notre méthode n'échappe pas à la règle. Globalement notre méthode a donné en moyenne des résultats intéressants lors de la campagne INEX 2004, puisque classée au 13^{e} rang sur 70 soumissions des participants pour les requêtes définissant uniquement le contenu textuel souhaité et au 5^{e} rang sur 51 soumissions pour les requêtes qui combinent contenu et références explicites à la structure XML (cf. chapitre 2, section 3.6.2.2). Cependant, en regardant les résultats obtenus pour chaque requête des différences d'efficacité existent. C'est pourquoi nous avons cherché à identifier les caractéristiques des requêtes pour lesquelles notre méthode donne des résultats satisfaisants et les caractéristiques des requêtes pour lesquelles notre méthode donne de moins bons résultats. Le but est de pouvoir définir des types de requête et ainsi trouver si une configuration de la méthode peut être adaptée à chaque type de requête. De plus, la définition de types de requête peut permettre l'identification des requêtes pour lesquelles notre méthode peut

8. http://www.inex.otago.ac.nz
9. http://trec.nist.gov

TREC3	Inq102 (1er)	...	Assctv2 (5e)	Optimal	SélectionP@5
Local	0,6259	...	0,4810	0,6259	0,6259
(5 premières	0,2699	...	0,3987	0,6115	0,5667
requêtes)	0,1806	...	0,3002	0,3803	0,2681
	0,7372	...	0,6164	0,7372	0,7372
	0,2504	...	0,3762	0,3762	0,2504
Global	0,4226	...	0,3539	0,4837	0,4593
				(+ 14,46 %)	(+ 8,68 %)
TREC5	ETHme1 (1er)	...	Cor5Mrf1 (5e)	Optimal	SélectionP@5
Local	0,0673	...	0,0773	0,2215	0,2215
(5 premières	0,0453	...	0,0396	0,0932	0,0453
requêtes)	0,6813	...	0,7971	0,8760	0,6813
	0,3262	...	0,2632	0,3262	0,3262
	0,1660	...	0,0534	0,1660	0,1660
Global	0,3165	...	0,2931	0,4128	0,3786
				(+ 30,43 %)	(+ 19,62 %)
TREC6	uwmt6a0 (1er)	...	LNmShort (5e)	Optimal	SélectionP@5
Local	0,3185	...	0,2127	0,4934	0,4753
(5 premières	0,7671	...	0,4686	0,7671	0,7671
requêtes)	0,6556	...	0,4750	0,6909	0,6556
	0,5000	...	0,0879	0,5000	0,2599
	0,0302	...	0,0227	0,0600	0,0600
Global	0,4631	...	0,2902	0,5217	0,4703
				(+ 12,65 %)	(+ 1,55 %)
TREC7	CLARIT98COMB (1er)	...	CLARIT98RANK (5e)	Optimal	SélectionP@5
Local	0,7112	...	0,7054	0,8468	0,7112
(5 premières	0,5068	...	0,4833	0,5068	0,5068
requêtes)	0,4281	...	0,3804	0,4546	0,4546
	0,1675	...	0,0866	0,4204	0,1675
	0,4555	...	0,4334	0,5429	0,5429
Global	0,3702	...	0,3351	0,4820	0,4067
				(+ 30,20 %)	(+ 9,86 %)

Tableau 4.8 – MAP locale et globale après fusion des cinq meilleurs systèmes (Hubert *et al.*, 2007a)

être appliquée et les requêtes pour lesquelles une autre approche peut être envisagée. La plupart des critères définis sont néanmoins transposables à d'autres contextes de RI.

Dans ces travaux (Englmeier *et al.*, 2006; Hubert et Mothe, 2006), nous étudions certaines caractéristiques des requêtes pour un ensemble de configurations de notre méthode de recherche. Nous tentons d'identifier des caractéristiques pour lesquelles notre approche est plus (respectivement moins) efficace notamment pour cibler dans quelle direction faire évoluer notre méthode. Nous comparons également plusieurs configurations pour tenter de mettre en évidence des différences d'efficacité pour un même type de requête et éventuellement pouvoir appliquer la configuration la plus efficace pour chaque type de requête. Dans cette même optique, nous étudions également d'autres approches pour estimer si les critères de distinction de requête sont également utilisables pour dégager les forces et faiblesses de chaque approche et éventuellement pouvoir les combiner pour améliorer les recherches d'information XML (RI XML).

5.1 Critères de caractérisation des requêtes

En analysant la définition des requêtes proposées lors des campagnes d'évaluation INEX, il est possible de recenser différents éléments :

- les termes,
- les groupes de mots,
- les préférences (ou préfixes),
- les contraintes structurelles.

Ces éléments conduisent directement à la définition de critères possibles pour caractériser les requêtes à savoir :

- le nombre de termes constituant une requête,
- la présence ou non de groupes de mots ou bien le nombre de groupes (les groupes de mots sont délimités par des guillemets dans l'expression de la requête),
- le nombre de mots simples,
- la présence ou non de préférences sur les mots voire le nombre de préférences positives et le nombre de préférences négatives,
- la présence ou non de contraintes de structure et plus précisément la présence de contraintes de granularité de résultat, la présence de contraintes sur les mots recherchés (contraintes de localisation) voire leur nombre. L'expression de contraintes de structure est complexe et peut engendrer une multitude de cas de figure. La définition de critères liés aux contraintes de structure réclame une étude approfondie qu'il reste à mener.

Parmi ces critères, nous avons étudié le nombre de termes et le nombre de groupes de mots. Cette liste est bien entendu extensible en prenant en compte par exemple des notions linguistiques vis-à-vis des termes constituant les requêtes (par exemple nombre de noms, verbes ou sigles). Il est aussi possible d'introduire un critère relatif à la présence ou au nombre de mots composés.

Au-delà des critères définis précédemment il est possible de définir des critères supplémentaires comme par exemple :

- le nombre d'unités constituant une requête c'est-à-dire le nombre de mots simples plus

le nombre de groupes de mots. Un utilisateur peut considérer que trouver une partie d'un groupe constitue déjà une certaine pertinence mais il peut également considérer un groupe de mot comme une unité à trouver intégralement,
– le nombre d'éléments de la collection contenant chaque terme de la requête et notamment le minimum des nombres d'éléments de la collection contenant un terme de la requête et le maximum des nombres d'éléments contenant un terme de la requête. Ce critère renseigne sur l'utilisation de termes plus ou moins discriminants dans la définition d'une requête.

Bien sûr, à partir de ces critères de base il est possible de définir des critères plus élaborés combinant plusieurs critères de base comme par exemple :

– la proportion de groupes de mots par rapport au nombre de termes ou par rapport au nombre d'unités si l'on suppose que l'utilisateur cherche à retrouver les unités qu'il a indiquées dans la définition de sa requête,
– l'écart entre le minimum des nombres d'éléments de la collection contenant un terme de la requête et le maximum des nombres d'éléments contenant un terme de la requête,
– la proportion de termes discriminants dans une requête,
– la proportion de termes ayant une préférence négative par rapport au nombre de termes.

Parmi ces derniers critères, nous avons étudié la proportion de groupes de mots par rapport au nombre d'unités.

Certains des critères mentionnés ont été utilisés dans le cadre de travaux d'intégration de techniques de traitement automatique de langues (TAL) en recherche d'information dont Moreau et Sébillot (2005) présentent un panorama. Le nombre de termes peut par exemple distinguer les requêtes pour lesquelles certaines techniques de TAL peuvent améliorer les performances de systèmes de recherche d'information.

5.2 Expérimentations relatives à l'étude du comportement du SRI suivant les caractéristiques des requêtes

Pour étudier le comportement de notre SRI XML face aux requêtes, nous avons réalisé des expérimentations en fonction de trois caractéristiques des requêtes : le nombre de termes constituant une requête, le nombre de groupes de mots constituant une requête et la proportion de groupes de mots par rapport au nombre d'unités. Les expérimentations accordent une grande place aux critères relatifs aux groupes de mots. Ceci est notamment dû au fait que notre méthode ne traitant pas de manière particulière les groupes de mots, ceux-ci peuvent donc constituer une limite. Des expérimentations complémentaires ont ensuite été réalisée afin de comparer le comportement d'autres SRI XML par rapport au notre SRI en fonction des même caractéristiques des requêtes.

5.2.1 Protocole d'expérimentation sur la collection IEEE – INEX 2004

Différentes expérimentations utilisant notre méthode ont été réalisées sur le jeu de requêtes de type CO utilisé durant la campagne d'évaluation INEX 2004 (Englmeier *et al.*, 2006; Hubert et Mothe, 2006). Ces expérimentations possèdent une configuration commune d'un ensemble de paramètres (méthode d'indexation, coefficients associés aux préférences, constante φ) de la mé-

thode résultant d'une phase d'entraînement réalisée sur les requêtes de type CO et CAS d'INEX 2003. La valeur de la constante φ a notamment été fixée à 400 (cf. section 3.1). L'efficacité de la méthode peut dépendre de l'indexation et ce paramètre est donc à étudier. Néanmoins, dans un premier temps, nous étudié l'influence de la couverture et de l'agrégation de score. Les configurations définies pour les expérimentations diffèrent donc sur les valeurs du seuil de couverture *CT* (cf. section 3.1) et du coefficient d'agrégation α (cf. section 3.1). Le libellé identifiant chaque expérimentation est construit de manière à identifier les valeurs choisies pour la couverture et l'agrégation : αX*CT*Y où X correspond à la valeur du coefficient de propagation de score α et Y correspond à la valeur du seuil de couverture vérifié. Par exemple, l'expérimentation libellée α0.25C0.35 est paramétrée avec une valeur de 0,25 pour le coefficient propagation de score α et avec une valeur du seuil de couverture *CT* égale à 0,35.

Il est à noter que plus le coefficient d'agrégation est grand, plus l'agrégation de score est faible. De plus, plus le coefficient de couverture est élevé plus la couverture de la requête doit être grande pour qu'un élément soit retenu.

5.2.2 Comportement de notre approche en fonction des caractéristiques des requêtes

Au regard des résultats représentés dans la figure 4.1, le premier constat est que le SRI semble plus efficace pour des requêtes constituées d'un nombre limité de termes (< 5) pour les configurations testées. Une explication peut-être que plus le nombre de termes augmente, plus une partie de la requête présente dans un élément peut suffire à le juger pertinent. Notre méthode semble alors limitée pour gérer partiellement des requêtes. Les résultats montrent également que la configuration peut influencer les résultats obtenus. Trop de couverture (*α0.05CT0.50*) dégrade les résultats lorsque le nombre de termes est plus important. Une couverture trop élevée supprime trop d'éléments pourtant apparemment jugés pertinents dans lesquels la requête est trop partiellement présente.

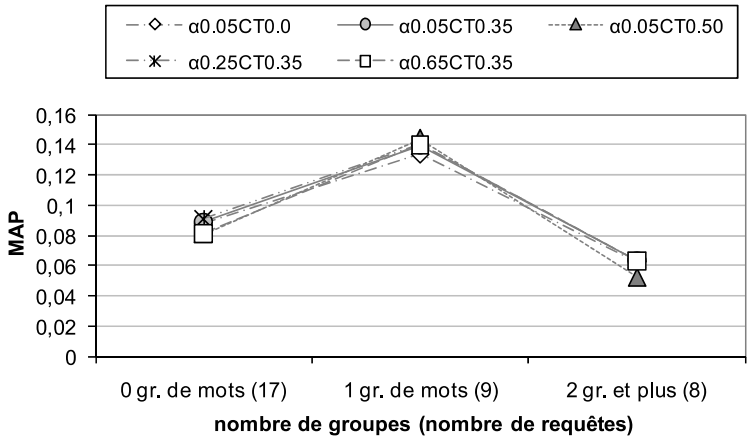


Figure 4.1 – Précision moyenne en fonction du nombre de termes

Les résultats synthétisés dans la figure 4.2 montrent que la méthode de recherche semble plus efficace pour des requêtes constituées d'un nombre limité de groupes de mots ($\leqslant 1$). Les résultats montrent également que les différentes combinaisons de propagation de score et de couverture donnent sensiblement les mêmes résultats (moins de 6 % de différence en moyenne). Les résul-

tats plus faibles lorsque le nombre de groupes de mots augmente peuvent s'expliquer par le fait que les groupes de mots ne sont pas traités de manière spécifique mais comme des mots simples indépendants.

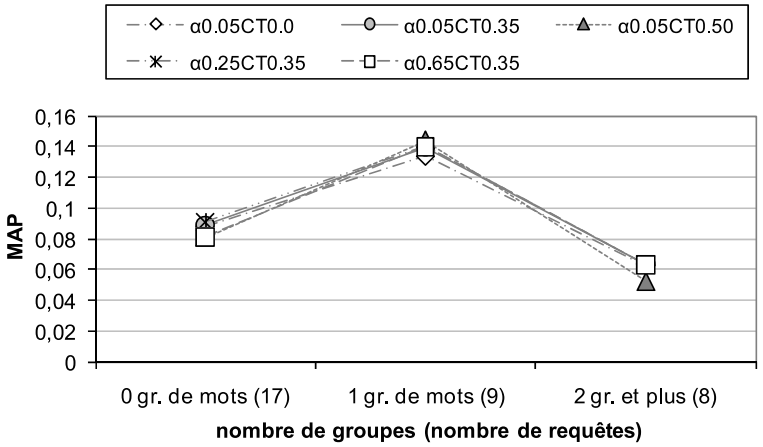


Figure 4.2 – Précision moyenne en fonction du nombre de groupes de mots

La figure 4.3 montre que la méthode de recherche semble plus efficace pour des requêtes pour lesquelles des groupes de mots sont présents et représentent jusqu'à 50 % des concepts recherchés. Au-delà de 50 % de groupes de mots, les résultats sont dégradés ce qui peut s'expliquer par le fait que notre méthode ne traite pas la recherche d'expression. Lorsqu'une requête est quasiment constituée de groupes de mots, il semble que la notion de groupe de mots soit davantage à respecter par rapport aux requêtes constituées d'une part plus importante de mots simples.

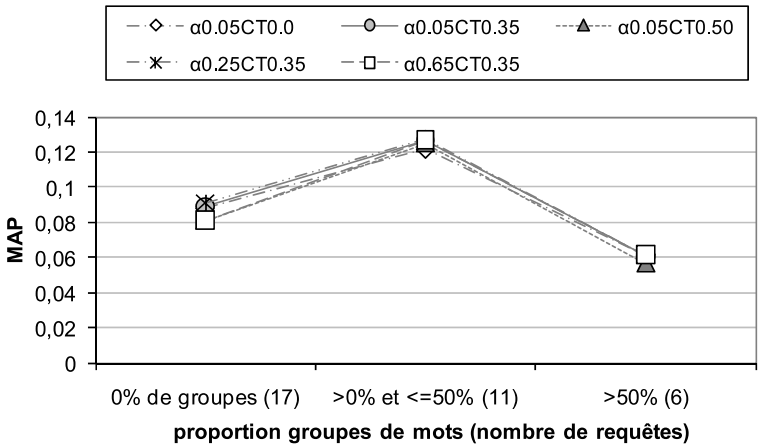


Figure 4.3 – Précision moyenne en fonction de la proportion de groupes de mots

5.2.3 Comparaison avec le comportement d'autres approches

En plus d'identifier des critères permettant de distinguer des différences de comportements de notre méthode suivant les requêtes, il nous a semblé intéressant de regarder si ces critères mettaient en évidence des comportements différents pour d'autres approches. Pour cela, nous avons choisi d'étudier les résultats ayant obtenu les trois meilleurs rangs en moyenne toutes requêtes et

mesures confondues globalement lors de la même campagne INEX 2004.

Le libellé identifiant chaque résultat (« run ») indique le nom du participant ayant soumis l'expérimentation lors de la campagne INEX 2004. Ces expérimentations présentent également plusieurs caractéristiques intéressantes :

- deux résultats (nommées IBM Haïfa1 et IBM Haïfa2) correspondent à des variantes d'une même approche (IBM Haïfa). Ceci permet de regarder si, comme observé pour notre méthode, les variantes d'une même approche présentent un comportement similaire au regard des critères étudiés,
- deux approches différentes sont basées sur des modèles différents. De plus ces approches sont également différentes de la notre. L'approche correspondant à IBM Haïfa1 et IBM Haïfa2 (classés respectivement 1er et 2e) est fondée sur le modèle vectoriel, l'utilisation de plusieurs index à différents niveaux et la mesure de similarité Cosinus (Mass et Mandelbrod, 2005). L'approche correspondant à U. Waterloo (classé 3e (Clarke et Tilker, 2005) est fondée sur un modèle probabiliste et la mesure Okapi BM25 (Robertson *et al.*, 1998). Ces différences permettent d'observer si des approches différentes présentent des différences de comportement vis-à-vis des critères étudiés. Ces différentes approches rejoignent néanmoins la notre dans le fait qu'elles ne traitent pas de manière particulière les groupes de mots.

Les données directement issues des systèmes choisis (IBM Haïfa et U. Waterloo) et soumises lors de la campagne INEX 2004 ont été utilisées pour réaliser les analyses présentées.

Au regard des résultats représentés en figure 4.4, nous constatons que :

- toutes les approches suivent la même tendance face au nombre de termes de la requête : lorsque le nombre de termes est important (7 termes ou plus), l'efficacité de la méthode se dégrade. Notre approche *α0.25CT0.35* voit ses résultats dégradés dès 5 termes ; ce sont ces requêtes de 5 ou 6 termes qui pénalisent notre approche,
- les requêtes de moins de 4 termes sont celles pour lesquelles les différences d'efficacité entre approches sont les plus importantes. C'est sur ces requêtes que le système d'IBM Haïfa montre une plus grande efficacité. Notre approche *α0.25CT0.35* qui se classerait globalement au 6e rang atteint le 3e rang pour ce type de requête,
- les variantes 1 et 2 de la même approche IBM Haïfa ont des comportements très similaires au regard du nombre de termes constituant les requêtes. Ceci rejoint le constat précédemment fait pour des variantes de notre approche (cf. section 5.2.2).

Les résultats présentés dans la figure 4.5, montrent des différences de comportement entre les trois approches. Les approches IBM Haïfa (1 ou 2) et U. Waterloo ont des comportements proches exceptés au niveau des requêtes comportant 2 groupes de mots ou plus. Notre approche présente un manque d'efficacité pour les requêtes constituées uniquement de mots simples ou constituées d'au moins 2 groupes de mots. Néanmoins, notre approche se distingue par une efficacité nettement plus importante (> 40 %) pour les requêtes constituées de seulement 1 groupe de mots.

La figure 4.6 montre que les trois expérimentations les mieux classées globalement lors d'INEX 2004 ont principalement une efficacité supérieure pour les requêtes ayant une proportion de groupes de mots nulle (50 % des requêtes). De même, les deux variantes du même système classées aux deux premiers rangs montrent des résultats très nettement supérieurs pour des requêtes constituées majoritairement de groupes de mots c'est-à-dire une proportion de groupes de mots

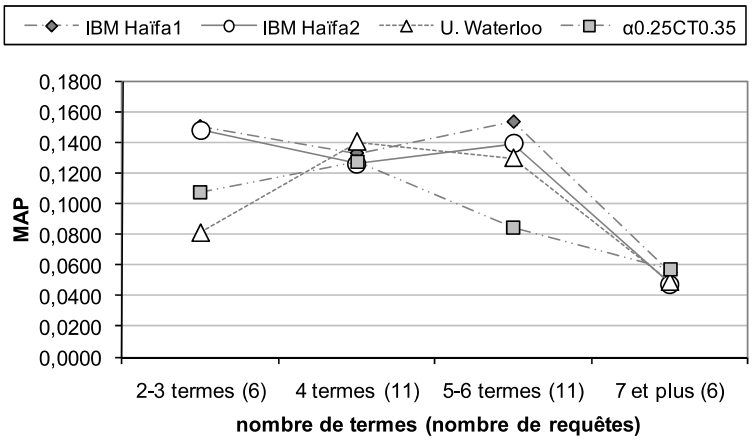


Figure 4.4 – Précision moyenne en fonction du nombre de termes

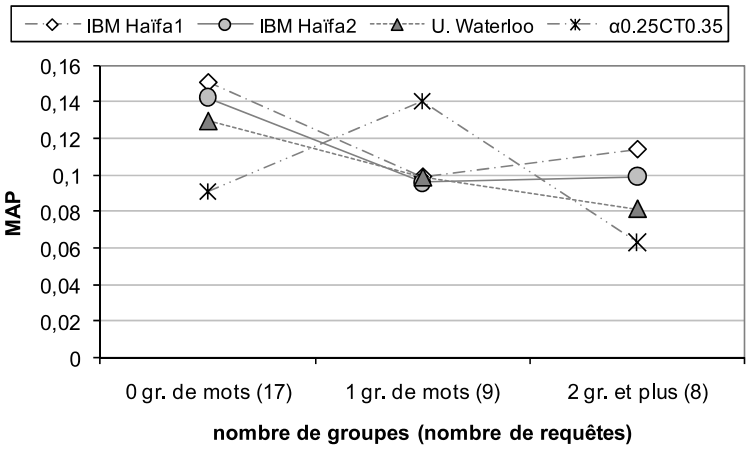


Figure 4.5 – Précision moyenne en fonction du nombre de groupes de mots

supérieure à 50 %. À l'inverse, notre approche montre une meilleure efficacité pour une proportion de groupe de mots inférieure ou égale à 50 % mais non nulle.

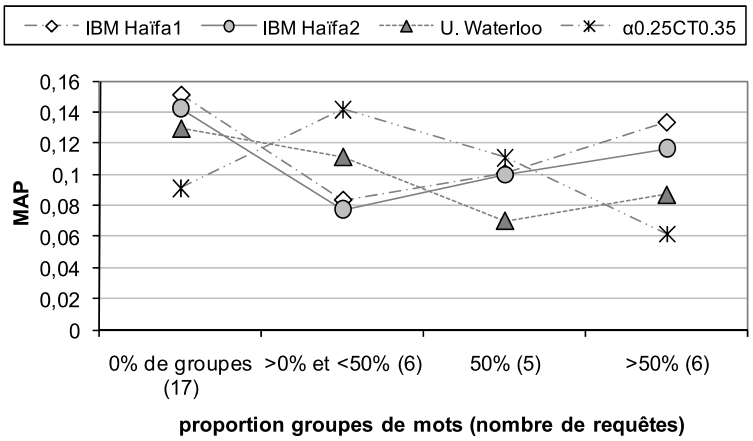


Figure 4.6 – Précision moyenne en fonction de la proportion de groupes de mots

5.2.4 Discussion

En analysant globalement les différents résultats obtenus et détaillés dans les différentes sections précédentes nous pouvons faire les observations suivantes :

- les critères sont utilisables pour différencier des comportements de notre SRI. Le nombre de termes de la requête met en évidence que la méthode dans certaines configurations produit les meilleurs résultats pour des requêtes ayant moins de 5 termes. Le critère de proportion de groupes de mots montre une faiblesse de la méthode pour les requêtes sans groupe de mot ou majoritairement composées de groupes de mots,
- les critères ne permettent pas réellement de dégager plusieurs configurations de la méthode qui soient efficaces pour des types de requêtes différents et donc qui pourraient être combinées pour améliorer les résultats. La meilleure configuration correspond à un coefficient de propagation de 0,25 et une couverture de 35 %. Il est donc nécessaire de poursuivre cette étude suivant d'autres critères, avec d'autres configurations impliquant d'autres paramètres,
- les critères permettent d'identifier des comportements similaires pour différentes approches. Par exemple, les requêtes longues (requêtes constituées d'un nombre important de termes) conduisent à des résultats plus faibles (en termes de précision moyenne),
- les critères permettent d'identifier des plages de valeurs pour lesquelles différentes approches fournissent des résultats variables. Par exemple, le nombre de termes montre une efficacité supérieure pour une approche (IBM Haïfa) pour les requêtes de peu de termes (moins de 4). Ces critères permettent également de pointer des points forts ou des faiblesses des approches. Par exemple, la proportion de groupes de mots montre que l'efficacité de notre méthode est supérieure pour une proportion inférieure ou égale à 50 % mais non nulle alors que d'autres approches sont supérieures lorsque la proportion est nulle ou au contraire supérieure à 50 % (IBM Haïfa),
- les critères permettent d'identifier les approches ayant des comportements différents vis-à-vis des critères et notamment des comportements complémentaires. Par exemple, notre approche semble davantage complémentaire à l'approche IBM Haïfa que l'approche U. Waterloo.

6 Choisir l'interface de restitution en fonction du scénario d'utilisation

La RI est généralement réalisée au travers d'une interface qui permet à l'utilisateur de spécifier sa requête et de visualiser les résultats de la recherche. Dans ce contexte, une attention particulière doit être portée sur la dernière étape du processus de recherche d'information *ad hoc* qui consiste à visualiser les résultats de recherche d'une manière spécifique au travers d'une interface visuelle de recherche d'information (VIRI). Ce dernier point est très important dans le processus de RI puisqu'il est censé faciliter le traitement des résultats de recherche par l'utilisateur (afin de trouver les documents pertinents pour lui). Un problème important est donc le choix judicieux d'une interface adaptée. Notre hypothèse est que ce choix est fonction du scénario de RI qui dépend de l'utilisateur effectuant la recherche, de la tâche qui motive sa recherche et du SRI sous-jacent qui traite la recherche. Le problème est donc de fournir une interface adaptée à chaque contexte de RI.

L'évaluation de VIRI est une tâche difficile qui reste à ce jour à compléter. Bien que des évaluations de VIRI existent, elles sont essentiellement axées sur l'étude de l'utilisabilité même si plusieurs niveaux d'évaluation peuvent être considérés (Thomas et Cook, 2005). Ce type d'évaluation est cependant incomplet de notre point de vue. Il ne permet pas, par exemple, d'évaluer si une VIRI donnée est réellement adaptée à un scénario de RI particulier (une tâche de RI donnée effectuée par un utilisateur donné sur un système donné).

De plus, pour les VIRI ayant fait l'objet d'évaluations, une évaluation spécifique à chaque VIRI a été menée ce qui rend difficile une réelle comparaison des VIRI entre elles (Julien *et al.*, 2008). Le problème dans l'évaluation de ces VIRI réside dans le fait qu'elles ne sont pas basées sur le même SRI et qu'elles ne proposent pas les mêmes traitements (comme par exemple une classification ou un filtrage de résultat). Ainsi, identifier l'interface « la plus adaptée » est plutôt difficile surtout quand ce jugement dépend de nombreux paramètres.

La définition d'un protocole d'évaluation de VIRI dans un cadre commun reste donc une problématique à laquelle nous proposons de répondre. Nous proposons d'évaluer l'adéquation de VIRI suivant différents scénarios de RI (Chevalier et Hubert, 2005; Bonnel *et al.*, 2008). Pour cela, l'évaluation se base, d'une part, sur une simulation de SRI. D'autre part, un ensemble de critères permettent de décrire différents scénarios de RI et de définir les jeux d'essai correspondant afin de tester si un utilisateur donné réalise une tâche de RI donnée à l'aide d'une VIRI donnée. Des techniques statistiques sont ensuite utilisées pour mesurer l'adéquation de chaque VIRI. Enfin, une application permet de trouver une VIRI adaptée en spécifiant un scénario de RI particulier.

6.1 Principes d'évaluation de l'adéquation d'une interface à un scénario de RI

Dans le cadre de la RI, évaluer l'utilisabilité d'une interface de restitution est nécessaire. Cependant, même si une interface possède une « bonne » utilisabilité, elle peut ne pas correspondre à un utilisateur donné réalisant une tâche particulière de recherche d'information. Il est donc nécessaire d'identifier quels sont les scénarios de RI où VIRI est appropriée est ceux où elle ne l'est pas. Pour cela, deux types d'évaluations peuvent être envisagées :

- L'évaluation d'une VIRI conjointement avec son propre SRI : Ce type d'évaluation s'inscrit dans des campagnes d'évaluations plus larges telles que TREC [10], CLEF [11] ou INEX [12]. Ces campagnes proposent une tâche spécifique qualifiée d'« interactive ». Cependant, les tâches de ces campagnes ne sont pas conçues pour évaluer la VIRI elle-même. Elles considèrent l'interface comme un composant d'un système de recherche d'information plus global. Même si ces tâches sont axées vers l'interaction avec l'utilisateur, l'influence de la VIRI dans le processus de RI est noyée dans la performance du SRI sous-jacent. Ainsi, cela suppose que l'influence de la VIRI est corrélée avec la performance du SRI. Suivant ce principe il est impossible d'identifier la part du SRI et la part de la VIRI dans l'efficacité du système ;
- L'évaluation d'une VIRI indépendamment de son SRI : Cette approche implique qu'une VIRI puisse être associée à n'importe quel SRI et puisse être considérée comme un composant. En fait, intuitivement nous pouvons penser qu'une VIRI peut améliorer la tâche de l'utilisateur. Par exemple, une VIRI basée sur des cartes auto-organisatrices peut l'identification des

10. Text REtrieval Conference – `http://trec.nist.gov`
11. Cross-Language Evaluation Forum – `http://clef.isti.cnr.it`
12. Initiative for the Evaluation of XML Retrieval – `http://inex.is.informatik.uni-duisburg.de`

documents pertinents dans un résultat fourni par un SRI avec une faible précision.

Le premier type d'évaluation présente l'inconvénient de la non séparation de l'évaluation de la VIRI et de son SRI. Nous considérons donc le second type d'évaluation (Chevalier et Hubert, 2005). Nous proposons une approche qui vise à évaluer une VIRI indépendamment du SRI pour identifier les contextes pour lesquels la VIRI est adaptée (c'est-à-dire pour lesquels l'utilisateur parvient à réaliser sa tâche avec succès). Cette approche s'appuie sur un SRI virtuel qui fournit des résultats de recherche à la VIRI évaluée. Ce SRI virtuel est basé sur des jeux d'essai spécifiques (requêtes + jugements de pertinence) qui simulent différents contextes RI suivant des critères précis. Ces critères, liés aux principales caractéristiques impliquées dans le processus de RI, sont introduits dans la section suivante.

6.2 Caractérisation des scénarios de RI

Pour évaluer l'adéquation d'une VIRI à différents scénarios de recherche, une première étape est de caractériser définir ces scénarios. Pour cela, nous modélisons un scénario de recherche comme un triplet (SRI, utilisateur, tâche) (Bonnel *et al.*, 2008). Nous définissons un scénario comme un ensemble de caractéristiques, chacune liée à un membre du triplet. En faisant varier la valeur de chaque caractéristique un scénario particulier est défini.

Différents critères sont identifiés dans (Lainé-Cruzel, 1999) : « Qui est l'utilisateur ? », « Que cherche-t-il ? », et « Quel est son but ? ». Hölscher et Strube (2000) confirment cette vision en caractérisant l'utilisateur suivant deux connaissances principales : connaissance du domaine et connaissance pratique.

Ainsi, dans notre approche, les critères sont séparés en trois catégories :

- ceux liés à l'utilisateur (« Qui est l'utilisateur ? »),
- ceux liés à la tâche de recherche d'information (« Que cherche-t-il ? », « Quel est son but ? »,
- ceux liés au système et aux résultats de recherche. Nous introduisons cette catégorie pour caractériser la difficulté que certains utilisateurs éprouvent lorsqu'ils manipulent certains types de résultats de recherche.

Les sections suivantes présentent différents critères, liés à chaque membre du triplet (SRI, utilisateur, tâche), qui peuvent être utilisées pour caractériser les scénarios de RI. Cette liste de critères n'est pas exhaustive, elle peut être étendue ou modifiée. En ajoutant des critères, des scénarios de RI plus précis peuvent être obtenus et en plus grand nombre. Cela offre un moyen plus fin de distinguer les VIRI du point de vue de l'adéquation.

6.2.1 Critères liés au SRI

Les critères liés au SRI caractérisent l'ensemble de documents restitué. En effet, le succès d'une tâche dépend en partie de ces critères. Par exemple, si l'on considère une tâche de recherche de type recherche d'un item connu, une liste ordonnée est supposée efficace si le SRI restitue l'item recherché dans les premiers rangs du résultat. Ces critères sont notamment liés à des notions habituellement utilisées pour évaluer les SRI, par exemple dans les campagnes TREC [13]. Les critères liés au SRI sont par exemple :

13. `http://trec.nist.gov`

- **Nombre de documents restitués.** Ce critère correspond au nombre de documents restitués par le SRI. En effet, une VIRI peut être efficace pour un nombre réduit de documents retournés, et au contraire devenir inefficace quand le nombre de document augmente.
- **Homogénéité de contenu des documents.** Ce critère mesure à quel point les contenus des documents retrouvés traitent du même sujet (d'un point de vue lexical et éventuellement conceptuel). Il permet, par exemple, d'évaluer si une VIRI est efficace lorsque de nombreux sujets sont abordés dans les documents retrouvés.
- **Précision.** Ce critère indique la proportion de documents pertinents par rapport au nombre de documents retrouvés par le SRI.
- **P@N.** Ce critère donne la proportion de documents pertinents parmi les N premiers documents retournés par le SRI.
- **Écart-type des rangs des documents pertinents.** Ce critère estime la répartition de documents pertinents dans le résultat retourné par le SRI. Il indique par exemple si les documents pertinents sont situés autour d'un rang donné dans le résultat.

6.2.2 Critères liés à l'utilisateur

Hölscher et Strube (2000) soulignent que le succès d'une recherche d'information dépend de deux critères principaux : le niveau de connaissance du domaine et le niveau de connaissance pratique. Nous nous sommes appuyés sur cette proposition pour définir les critères primaires liés à l'utilisateur :

- **Niveau de connaissance du domaine.** Ce niveau influence, entre autres, la manière dont l'utilisateur appréhende les documents du résultat. En effet, plus la connaissance du domaine est élevée, meilleure est l'identification des documents pertinents par l'utilisateur. L'utilisateur doit connaître le domaine correspondant à son besoin d'information pour identifier l'information pertinente.
- **Niveau de connaissance pratique.** Ce critère donne une information sur la capacité de l'utilisateur à utiliser un système de type SRI. Ce niveau de connaissance influence également la capacité de l'utilisateur à trouver les documents pertinents parmi l'ensemble des documents du résultat. Cette connaissance rassemble plusieurs notions comme l'utilisation d'un ordinateur, la manipulation d'un SRI et la manipulation d'interface graphique.

6.2.3 Critères liés à la tâche de recherche

Les critères relatifs aux tâches de recherche visent à identifier quel type de recherche peut être réalisé par l'utilisateur et quel est le degré de réalisation atteint par l'utilisateur.

Type de tâche. Différents types de tâches peuvent être associés aux besoins d'information des utilisateurs. Suivant la tâche, la VIRI ne restitue pas nécessairement le même résultat. Par exemple, Rosenfeld et Morville (1998) identifient les tâches de RI suivantes :

- **Recherche d'un item connu.** Pour cette tâche l'utilisateur connaît un document particulier. L'objectif n'est pas de retrouver des documents de manière exhaustive correspondant à un besoin mais de trouver le document escompté (Ogilvie et Callan, 2003).
- **Recherche d'existence.** Pour cette tâche l'utilisateur connait un sujet particulier. L'objectif n'est pas de retrouver des documents de manière exhaustive correspondant au besoin mais

de trouver au moins un document traitant du sujet.
- **Recherche exploratoire.** Pour cette tâche, l'utilisateur effectue une exploration autour d'un sujet, l'objectif n'étant pas de retrouver des documents de manière exhaustive correspondant au besoin mais d'en retrouver suffisamment.
- **Recherche complète.** L'objectif de cette tâche est de trouver de manière exhaustive les documents correspondant au besoin.

Comme indiqué précédemment la liste des critères décrivant les scénarios de RI n'est pas limité aux critères et de nombreux autres peuvent être ajoutés. Par exemple, d'autres critères liés à l'utilisateur peuvent être correspondre à des caractéristiques physiques tels que l'âge, le sexe, à des déficiences visuelles, à des caractéristiques comme le niveau de lecture ou les langues connues (McCracken *et al.*, 2003) ou à des stratégies de recherche (Belkin *et al.*, 1996).

Les critères caractérisant les scénarios étant déterminés, l'étape suivante est de les utiliser dans un cadre d'évaluation. La section suivante présente le cadre d'évaluation que nous proposons.

6.3 Cadre d'évaluation

Le cadre d'évaluation que nous proposons repose sur un SRI indépendant des VIRI évaluées. En effet, chaque VIRI doit être indépendant de son SRI (c'est-à-dire des données) pour éviter le biais d'une évaluation de la VIRI et de son SRI comme un tout. Selon Hearst (2009, Chap. 2), pour comparer deux VIRI le processus d'évaluation doit utiliser la même collection de documents. Notre proposition dépasse cette recommandation en utilisant également le même SRI pour évaluer les VIRI suivant la même référence. Le cadre proposé suit des directives similaires à celles décrites dans (Hearst, 2009, Chap. 2).

Afin d'évaluer l'adéquation de différentes VIRI à des scénarios de RI, chaque VIRI doit être indépendante du SRI (c'est-à-dire des listes de documents). Selon Hearst (2009, Chap. 2) Chaque VIRI doit donc pouvoir être combinée à d'autres SRI. Un cadre d'évaluation doit être constitué de différentes parties :

- des critères d'évaluation. Les critères d'évaluation permettent d'évaluer la capacité d'une VIRI à réaliser un test donné.
- des données de test (jeux d'essai). Les données de test décrivent les différents scénarios de RI suivant les critères liés au SRI, la tâche ainsi qu'un utilisateur souhaité pour l'interaction avec la VIRI évaluée.
- des résultats d'évaluation. Les résultats correspondent à un ensemble de tests. Chaque test rassemble une valeur pour chacun des critères d'évaluation pour un jeu d'essai donné (scénario).
- des analyses des résultats. Les analyses reposent sur des méthodes qui exploitent les résultats.

Ces différentes parties sont présentées dans les sections suivantes.

6.3.1 Critères d'évaluation

Pour mesurer l'adéquation d'une VIRI à un scénario, nous devons des critères liés à la réalisation de la tâche. Dans notre proposition actuelle, un seul critère est utilisé qui correspond à l'achèvement ou l'échec de la tâche.

Réalisation de la tâche. Ce critère donne l'information sur le degré de réalisation de la tâche. L'intérêt de ce critère est par exemple de déterminer quelles combinaisons (système, utilisateur, tâche) sont adaptées à une VIRI donnée et lesquelles ne le sont pas. Pour ce critère, l'utilisateur indique s'il estime avoir réussi ou non à réaliser la tâche de recherche.

Lié à la réalisation la tâche, un critère complémentaire peut être le temps passé à réaliser la tâche. Il est cependant nécessaire dans ce cas de s'assurer que l'utilisateur utilise ce temps exclusivement pour l'évaluation.

6.3.2 Jeux d'essai

Pour être comparées selon la même base, les VIRI doivent utiliser les mêmes ensembles de résultats : les documents restitués par le SRI doivent être identiques pour toutes les VIRI. En effet, ce principe garantit la cohérence des résultats entre VIRI, pour assurer que chaque interface est évaluée de la même manière. Ceci n'exclut pas que chaque VIRI exploite différemment les documents. Cependant, pour permettre une telle exploitation, le SRI doit fournir suffisamment d'information sur chaque document (mots-clés, contenu, liens...).

De plus, une attention particulière soit être accordée à la variété des résultats de recherche utilisés (restitués par le SRI) durant le processus d'évaluation. Pour varier les scénarios, les résultats de recherche couvrir toutes les combinaisons des critères liés au système (nombre faible ou élevé de documents, faible ou haute précision, ...). Construire de tels jeux de test est une tâche difficile et très coûteuse en temps. Pour limiter ces inconvénients, les jeux d'essai peuvent s'appuyer sur des jeux existants comme ceux de TREC ou CLEF par exemple. Dans ce contexte, une première étape consiste à sélectionner ou ajouter des besoins en information (« topics ») pour couvrir les différentes tâches ; la seconde étape consiste à adapter les résultats de recherche (basés sur les jugements de pertinence ou *qrels*[14]) de manière à couvrir toutes les combinaisons de critères liés au SRI possibles (par exemple, en supprimant, ajoutant, modifiant le classement des documents dans le résultat). Pour réaliser cette étape, nous avons implémenté une application spécifique qui génère un résultat de recherche correspondant à toute combinaison des critères liés au SRI.

Les utilisateurs qui participent aux évaluations doivent également être soigneusement choisis de manière à construire un panel représentatif d'utilisateurs et donc un ensemble suffisant de scénarios. Ce panel doit couvrir les différentes combinaisons de critères liés à l'utilisateur (connaissance du domaine et connaissance pratique). Les évaluateurs doivent donc être répartis dans les 9 classes possibles et en nombre suffisant pour chaque classe afin de pouvoir établir des conclusions.

14. « qrels » est une liste des documents jugés pertinents pour un besoin en information donné et éventuellement des documents jugés non pertinents et non jugés.

6.3.3 Résultats d'évaluation

Les jeux d'essai présentés précédemment permettent d'évaluer plusieurs situations d'utilisation de VIRI. Il est alors possible de rassembler les résultats d'évaluation dans un tableau récapitulatif dont un exemple est donné dans le tableau 4.9.

ID éval	Scénario								Résultat
	SRI (documents restitués par le SRI)					Utilisateur (Niveaux de connaissances)		Tâche (But)	
	NbDocs	HCont.	Préc	P@N	ÉcTRang	Domaine	Pratique	Type	
1	24	élevé	0,75	0,9	2,3	débutant	débutant	connu	oui
2	250	faible	0,60	0,2	4,0	débutant	débutant	connu	non
…	…	…	…	…	…	…	…	…	…

Tableau 4.9 – Exemple de résultats d'évaluation

La section suivante présente comment les résultats ainsi collectés peuvent être analysés.

6.3.4 Exploitation des résultats d'évaluation

Pour caractériser précisément chaque VIRI évaluée, nous proposons, par exemple, d'analyser les résultats pour obtenir un arbre de décision. Un tel arbre permet notamment d'identifier les contextes pour lesquels la VIRI est efficace. Pour construire cet arbre de décision, l'algorithme C4.5 (Quinlan, 1993) peut par exemple être appliqué.

La figure 4.7 illustre un exemple d'arbre de décision résultant de l'analyse des résultats de l'évaluation d'une VIRI basée sur une visualisation sous forme de liste ordonnée de documents. Dans cet arbre de décision, nous constatons que la VIRI basée sur des listes ordonnées est efficace, par exemple, quand le nombre de documents pertinents est élevé dans le résultat de recherche et pour une tâche de type item connu.

6.4 Mise en œuvre du cadre d'évaluation

Nous développons une plateforme logicielle qui supporte le cadre d'évaluation de VIRI proposé. Cette plateforme facilite la réalisation d'expérimentations et leur reproductibilité. Une vue globale de l'environnement d'évaluation est synthétisée en figure 4.8.

Le prototype de plateforme est développé en Java. L'environnement d'évaluation repose sur un SRI virtuel nommé *VSE* (« virtual search engine ») partagé par toutes les VIRI évaluées. Pour permettre la communication entre la VIRI évaluée et l'environnement d'évaluation, le *VSE* fournit un ensemble de services relatifs notamment à la connexion, aux documents et requêtes manipulés.

Le *VSE* repose sur une base de données qui se décompose en trois catégories de données :

– les données relatives aux utilisateurs, aux VIRI et aux résultats d'évaluation correspondant,
– les données relatives aux critères impliqués dans le processus d'évaluation,

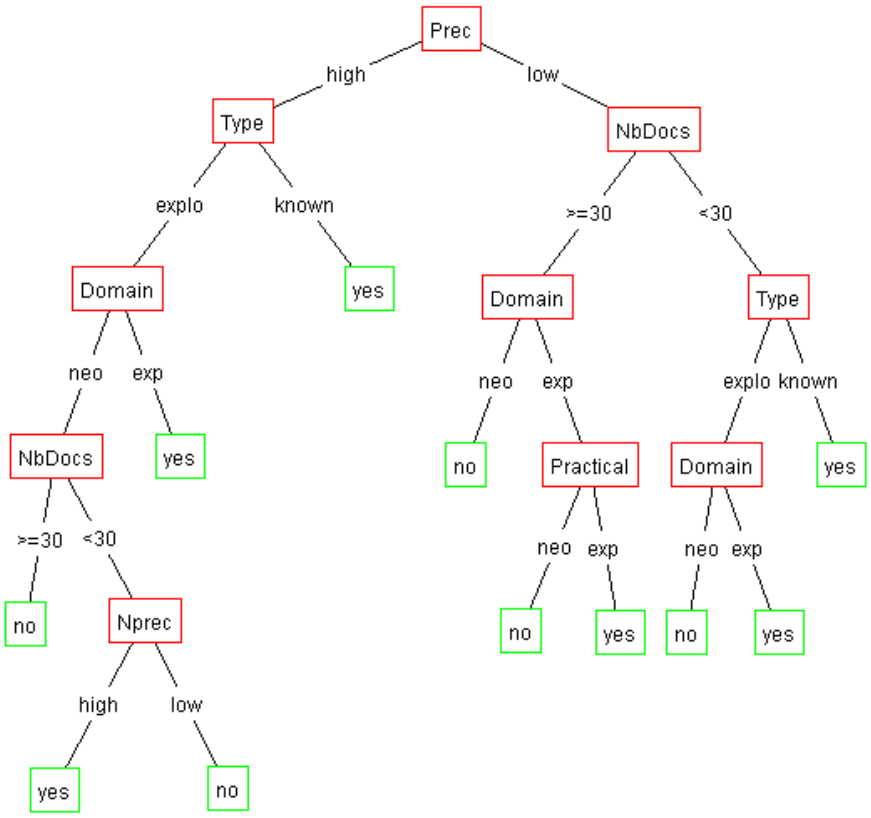


Figure 4.7 – Exemple d'arbre de décision résultant de l'évaluation d'une VIRI basée sur des listes ordonnées

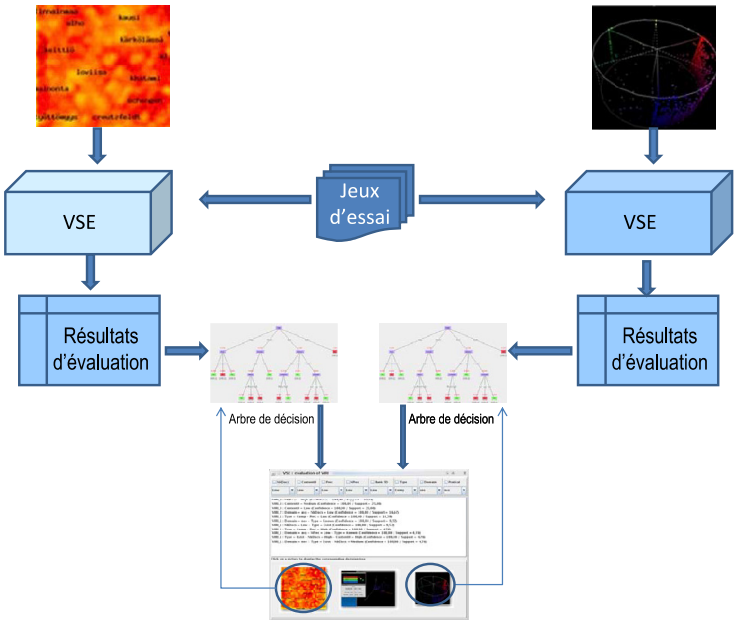


Figure 4.8 – Vue générale de l'environnement d'évaluation

– les données relatives aux jeux de test utilisés pour l'évaluation.

Le déroulement du processus d'évaluation est illustré par les figures suivantes. Au travers

d'une fenêtre de connexion, les utilisateurs sélectionnent son rôle (évaluateur, développeur de VIRI ou analyste). Un analyste a accès à tous les résultats de toutes les évaluations tandis qu'un développeur peut accéder seulement correspondant à sa VIRI. Une fois connecté, l'évaluateur doit choisir l'évaluation qu'il souhaite réaliser (cf. figure 4.9). Après le choix d'une évaluation donnée (requête), il indique son niveau de connaissance dans le domaine correspondant à l'évaluation.

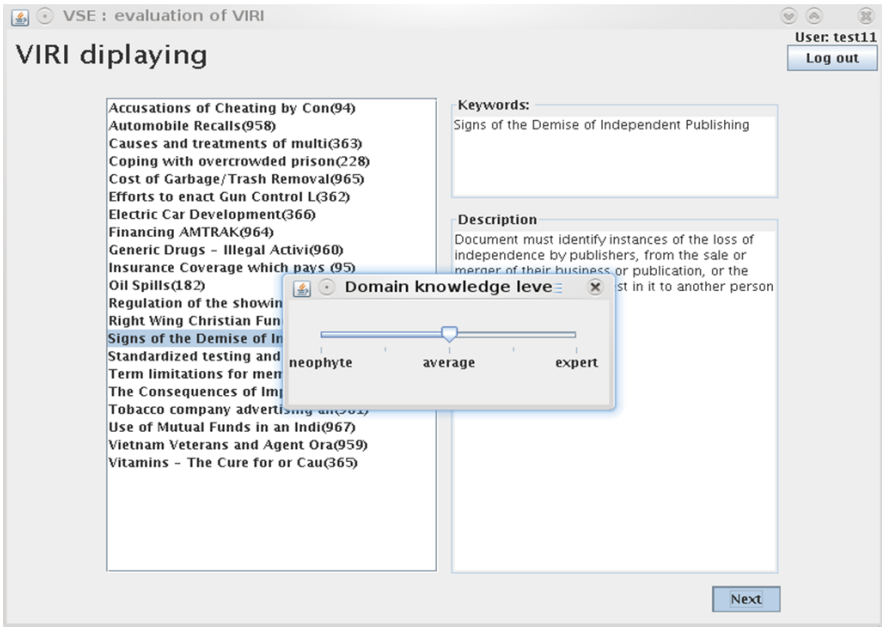


Figure 4.9 – Fenêtre de sélection des requêtes

Le *VSE* affiche ensuite l'objectif de l'évaluation à l'évaluateur (cf. figure 4.10). Il explique les résultats de recherche attendus pour réaliser la tâche (par exemple, recherche d'item connu).

Figure 4.10 – Explication relative à l'objectif de l'évaluation

En explorant les résultats au travers de la VIRI évaluée (par exemple, la liste ordonnée dans la fenêtre de gauche de la figure 4.11), l'utilisateur peut accéder au contenu d'un document particulier dans une fenêtre séparée (cf. fenêtre de droite de la figure 4.11) suite à l'appel de la VIRI au service du *VSE* correspondant.

De plus, l'évaluateur doit sélectionner les documents qu'il considère pertinents par rapport à la tâche qui lui est assignée. Pour cela, il peut soit cliquer sur le bouton « Ajouter à la liste de documents pertinents » dans la fenêtre affichant le contenu du document (cf. figure 4.11), soit sélectionner les documents dans la fenêtre d'évaluation dédiée ouverte par le *VSE* (cf. figure 4.12). Dans ce dernier cas, l'utilisateur sélectionne le document et l'ajoute à la liste des documents pertinents ordonnée par pertinence (cf. partie droite de la figure 4.12).

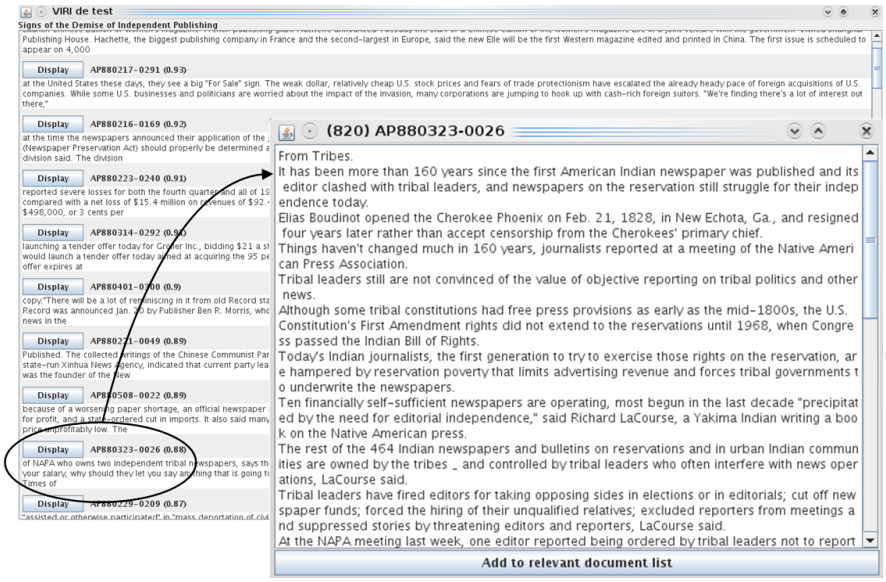


Figure 4.11 – VIRI particulière (fenêtre de gauche) et affichage du contenu d'un document fourni par un service du *VSE*

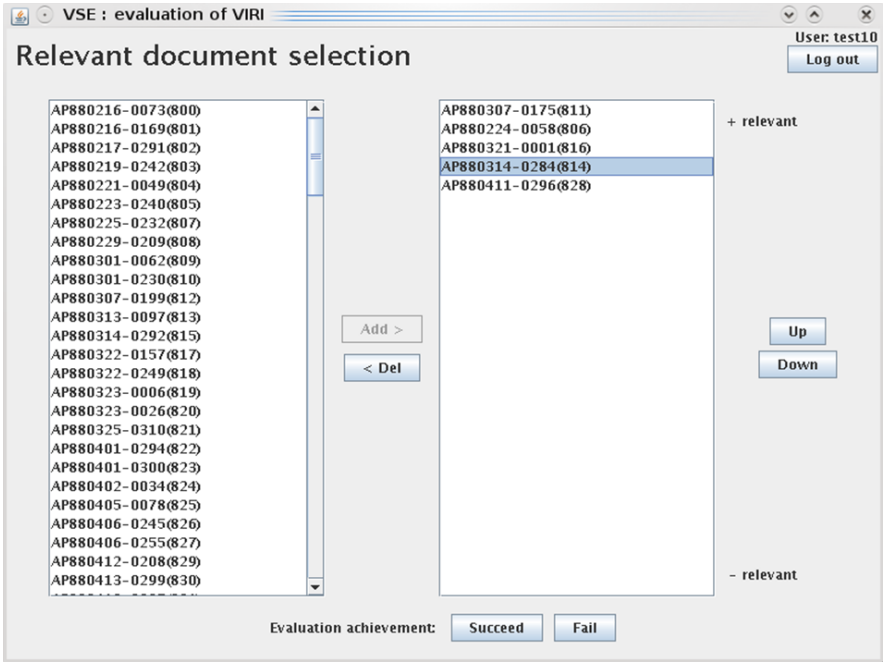
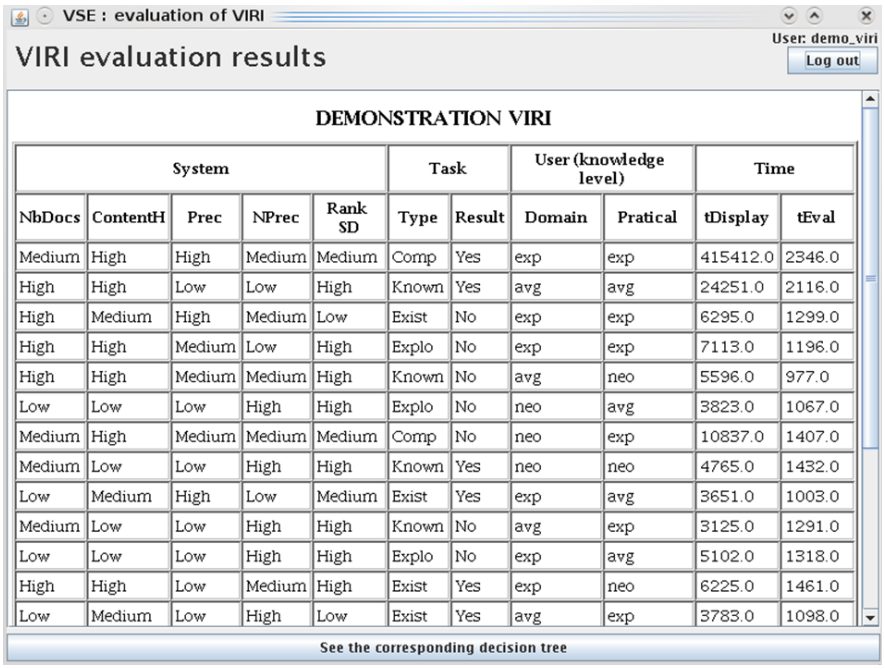


Figure 4.12 – Fenêtre de classement des documents jugés pertinents

Enfin, pour conclure son évaluation, l'utilisateur doit indiquer s'il considère qu'il a réalisé sa tâche ou s'il n'y est pas parvenu (cf. bas de la figure 4.12). Les résultats de l'évaluation (classement des documents par l'utilisateur et réalisation de la tâche) sont ensuite stockés dans la base de données.

Les résultats des évaluations étant stockés en lien avec chaque VIRI, chaque développeur de VIRI peut avoir accès aux résultats concernant sa VIRI (cf. figure 4.13). En complément du tableau récapitulatif des résultats, le développeur a également accès à l'arbre de décision (cf. figure 4.14)

basé sur les valeurs d'achèvement des tâches. Chaque développeur peut donc voir quels sont les scénarios pour lesquels sa VIRI est efficace et éventuellement travailler sur les scénarios pour lesquelles celle-ci n'est pas efficace.



System					Task		User (knowledge level)		Time	
NbDocs	ContentH	Prec	NPrec	Rank SD	Type	Result	Domain	Pratical	tDisplay	tEval
Medium	High	High	Medium	Medium	Comp	Yes	exp	exp	415412.0	2346.0
High	High	Low	Low	High	Known	Yes	avg	avg	24251.0	2116.0
High	Medium	High	Medium	Low	Exist	No	exp	exp	6295.0	1299.0
High	High	Medium	Low	High	Explo	No	exp	exp	7113.0	1196.0
High	High	Medium	Medium	High	Known	No	avg	neo	5596.0	977.0
Low	Low	Low	High	High	Explo	No	neo	avg	3823.0	1067.0
Medium	High	Medium	Medium	Medium	Comp	No	neo	exp	10837.0	1407.0
Medium	Low	Low	High	High	Known	Yes	neo	neo	4765.0	1432.0
Low	Medium	High	Low	Medium	Exist	Yes	exp	avg	3651.0	1003.0
Medium	Low	Low	High	High	Known	No	avg	exp	3125.0	1291.0
Low	Low	Low	High	High	Explo	No	exp	avg	5102.0	1318.0
High	High	Low	Medium	High	Exist	Yes	exp	neo	6225.0	1461.0
Low	Medium	Low	High	Low	Exist	Yes	avg	exp	3783.0	1098.0

Figure 4.13 – Tableau rassemblant les résultats d'évaluation d'une VIRI particulière

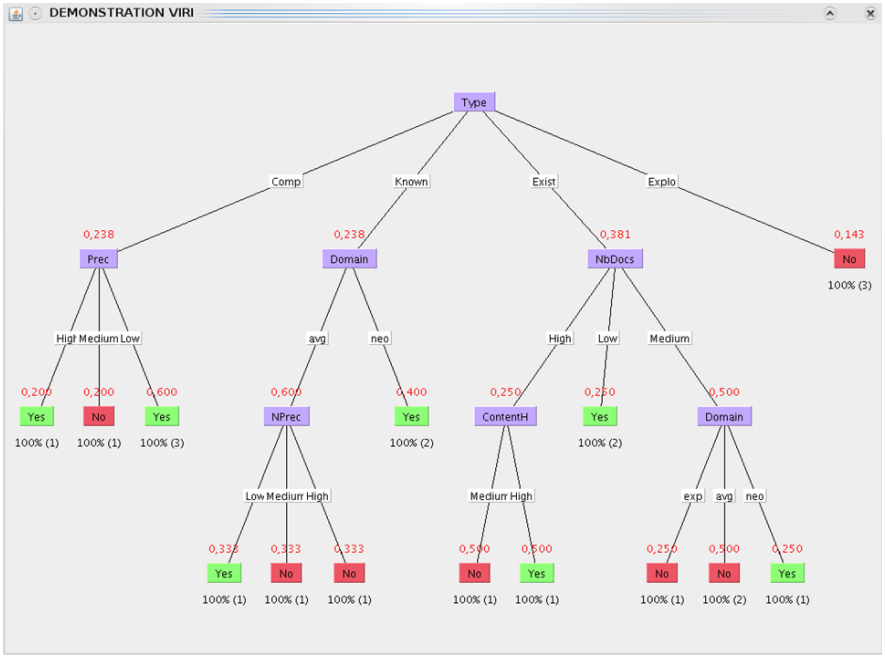


Figure 4.14 – Arbre de décision représentant le comportement d'une VIRI particulière

Les arbres de décision caractérisent le comportement de chaque VIRI évaluée séparément. Pour agréger plusieurs résultats afin de classer des VIRI par groupe, toutes les règles conduisant à la réalisation avec succès sont extraites des arbres de décisions. Ensuite, toutes ces règles sont af-

fichées dans une interface spécifique comme illustré dans la figure 4.15. Cette interface graphique est composée de trois parties (cf. figure 4.15) :

- la partie haute permet de construire un scénario de RI à partir des différents critères. Les cases à cocher (chacune correspondant à un critère) sélectionnées définissent le scénario de RI à prendre en compte et sont utilisées pour filtrer l'ensemble des règles (cf. figure 4.16) ;
- la partie centrale affiche les règles disponibles correspondant au scénario spécifié et la VIRI correspondante ;
- la partie basse affiche les miniatures des VIRI évaluées pour lesquelles les règles satisfont le scénario défini. En cliquant sur une miniature, l'arbre de décision représentant le comportement de la VIRI est affiché.

Ainsi, chaque utilisateur peut définir interactivement le contexte qui l'intéresse et pour lequel il espère trouver une VIRI adaptée.

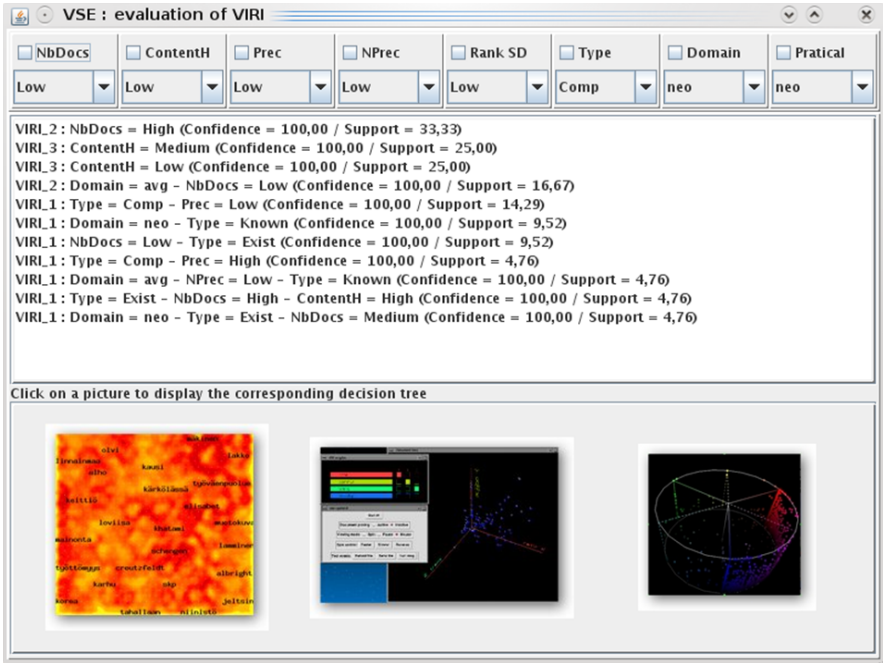


Figure 4.15 – Outil d'exploitation des résultats d'évaluations

7 Bilan et perspectives

Dans ce chapitre, nous avons détaillé nos contributions concernant la recherche de l'adéquation entre les outils de recherche et les recherches effectuées par les utilisateurs. L'objectif est, à l'inverse de la majorité des outils qui proposent le même comportement et le même traitement quelle que soit la demande de recherche, d'offrir la possibilité d'adapter le traitement à chaque recherche.

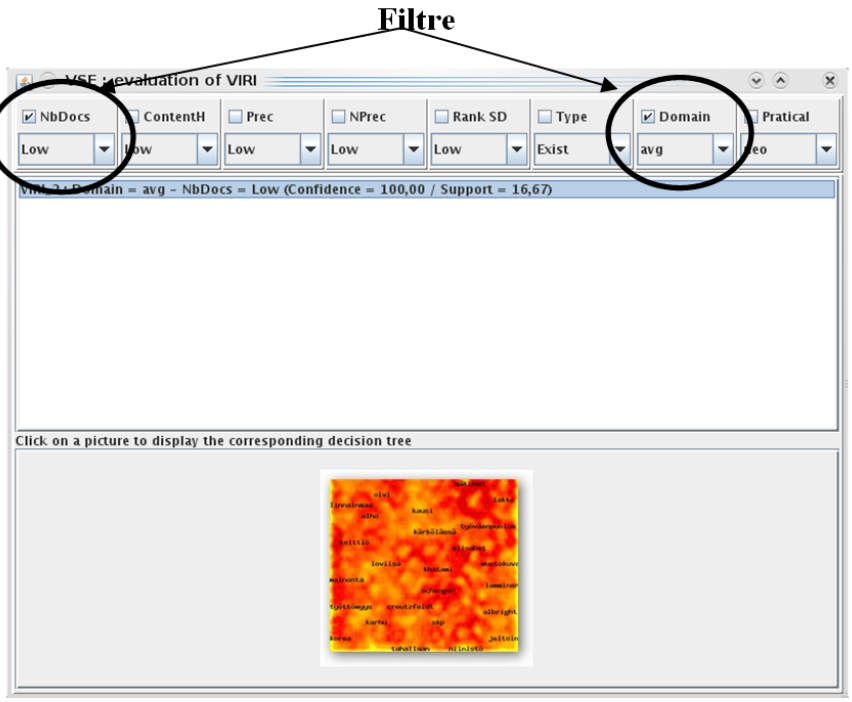


Figure 4.16 – Filtre de sélection d'interfaces

7.1 Contributions

Dans le cadre de la recherche de l'adéquation entre les outils de recherche et les recherches effectuées par les utilisateurs nos contributions ont suivi plusieurs orientations :

- une première orientation s'intéresse à la sélection du système efficace répondant à des éléments contextuels donnés. Dans cette orientation, une première proposition a été d'identifier comment agir sur les paramètres de notre SRI XML pour obtenir une meilleure précision des résultats en fonction de préférences de l'utilisateur. Une deuxième contribution a été d'étudier les méthodes de combinaison de résultats issus de plusieurs SRI. Nous avons notamment défini une méthode de combinaison par sélection du SRI le plus efficace en se basant sur le retour de pertinence fourni par l'utilisateur sur les premiers documents restitués par différents SRI. Ces travaux s'appuient sur des expérimentations réalisées sur des collections d'évaluations de référence : les collections INEX dans le domaine de la RI XML et les collections TREC dans le domaine de la RI non structurés.
- une seconde orientation s'intéresse à l'identification des contextes pour lesquels un système est efficace et ceux pour lesquels il ne l'est pas. Dans cette orientation, dans le domaine de la RI XML, nous avons identifié des critères de distinction des requêtes et leur influence sur l'efficacité de notre SRI XML. Ces travaux s'appuient sur des expérimentations réalisées sur les collections INEX.

  Par ailleurs, nous nous sommes intéressés également à l'évaluation des VIRI qui constituent un composant incontournable du processus d'évaluation et qui peut être adapté à chaque scénario de RI. Pour cela, nous avons défini un cadre d'évaluation de l'adéquation entre une VIRI et différents scénarios de RI. Il est ensuite possible d'adapter un système en sélectionnant la VIRI adaptée à un scénario de RI donné. Cette proposition a été mise en œuvre au

sein d'un prototype de plateforme d'évaluation baptisé *VSE*.

Ces travaux ont été menés en lien avec le projet européen Ws-Talk, le projet Égide Alliance Grande-Bretagne QUEST et le projet ACRIC du programme pluri-formations FRÉMIT.

7.2 Encadrement et diffusion scientifique

Les travaux menés dans cet axe ont donné lieu à différentes publications :

Thème	Publications
Sélection du système efficace pour un contexte	Conférences internationales : – ICEIS'07 (Hubert et Mothe, 2007a) – WEBIST'07 (Hubert *et al.*, 2007c) Conférences nationales : – VSST'07 (Hubert *et al.*, 2007a) – CIDE'07 (Hubert *et al.*, 2007b)
Identification des contextes pour lesquels un système est efficace	Revue nationale : – I^3 (Hubert et Mothe, 2006) Chapitre d'ouvrage national : – Information & visualisation - enjeux, recherches et applications (Bonnel *et al.*, 2008) Conférences et ateliers nationaux : – CORIA'06 (Englmeier *et al.*, 2006) – EGC'05 (Chevalier et Hubert, 2005)

7.3 Perspectives

Les perspectives à ces travaux visent, premièrement, à poursuivre les travaux sur la combinaison de SRI. Il sera intéressant de proposer une approche qui applique une combinaison différente en fonction des requêtes. L'objectif est de pouvoir appliquer la combinaison la plus performante pour chaque requête soumise. Nos contributions relatives à la caractérisation des requêtes qui doivent être poursuivies serviront également à mener ensuite cette approche.

Une seconde perspective est également de proposer une approche de type « Learning to rank » basée sur notre modèle de RI. En effet, différents paramètres peuvent varier et différentes fonctions peuvent être remplacées. Nos propositions se sont basées, pour l'instant, sur un apprentissage manuel, une perspective naturelle est d'automatiser ce processus.

Enfin, une troisième perspective concerne l'évaluation des VIRI. Un prolongement direct à nos travaux est de mettre en place une évaluation de différentes VIRI à l'aide de notre plateforme pour valider nos propositions et obtenir une première série de résultats. Ces travaux aboutiront à la définition d'une campagne d'évaluation de VIRI. Ils visent également à proposer un système qui s'adapte au scénario de RI effectué en fournissant la VIRI appropriée.

Conclusion générale

1 Synthèse

Mes travaux de recherche visent à prendre en compte les éléments de contexte dans le processus de RI. Dans cette optique, les problématiques générales auxquelles nous avons proposé des solutions ont été : comment intégrer les éléments de contexte et comment les exploiter pour adapter le système au contexte. Les éléments de contexte considérés ont été : le domaine, la structure des documents et l'utilisateur. Mes travaux se répartissent donc suivant quatre orientations :

- prise en compte du domaine dans la RI,
- prise en compte de la structure des documents,
- prise en compte de l'utilisateur,
- adaptation du comportement du système.

1.1 Première orientation : Prise en compte du domaine

Suivant cette orientation, notre première contribution a consisté à définir un modèle de RI qui supporte une RI basée sur une description du domaine sous forme de hiérarchies de concepts. Ce modèle est applicable à la recherche de concepts pour décrire les documents (processus de catégorisation) et, inversement, est applicable à la recherche de documents à partir de concepts. Il est également applicable à une RI *ad hoc* en texte libre et offre de surcroît la possibilité de combiner recherche en texte libre et recherche basée par concepts. L'intérêt de cette combinaison répond notamment aux utilisateurs qui ne parviendraient pas à identifier des concepts décrivant la totalité de leur besoin d'information. L'efficacité des différentes propositions a été évaluée au travers d'expérimentations et comparativement à des solutions populaires de la littérature. Ces propositions ont été mises en œuvre dans le cadre des projets européens IRAIA et eStage. Ce modèle sert également de base à des propositions suivant nos autres orientations.

La seconde contribution s'est intéressé à la RI sémantique à partir d'ontologie au travers de la proposition d'un modèle d'indexation sémantique. Le modèle vise à prendre en compte la dynamique de la collection (ajout, suppression ou modification d'un document). Le projet ANR DynamO constitue un cadre applicatif à nos travaux en RI sémantique.

1.2 Seconde orientation : Prise en compte de la structure des documents

Cette orientation a été menée pour répondre aux problématiques introduites lorsque la structure des documents est explicitement décrite et manipulable. La RI XML est un domaine qui rassemble ces problématiques. Nos contributions ont été basées sur le modèle de RI précédemment défini dans le cadre de notre première orientation pour des documents non structurés (cf. section 1.1). Nous avons étendu notre modèle au niveau de l'indexation des documents et des requêtes, et au niveau de l'appariement entre documents et requêtes pour intégrer les aspects liés à la structure des documents XML. L'indexation des documents tient compte des niveaux de granularité des éléments constituant un document XML. L'indexation des requêtes prend en compte les contraintes de structure relatives à la localisation d'un terme souhaitée par l'utilisateur et à la granularité souhaitée pour les éléments restitués. L'appariement entre requêtes et documents prend en compte de manière incrémentale les contraintes de structure et l'organisation hiérarchique des éléments XML au travers d'un principe d'agrégation de score.

Les propositions ont été menées dans le cadre du projet européen Ws-Talk et du projet Égide Alliance Grande-Bretagne QUEST. Elles ont été mises en œuvre au sein d'un prototype de SRI XML. Ces travaux nous ont conduits à participer à quatre éditions successives du programme d'évaluation INEX. Ces participations nous ont permis de mener différentes expérimentations et d'évaluer l'efficacité des différentes propositions et leurs limites.

1.3 Troisième orientation : Prise en compte de l'utilisateur

Suivant cette troisième orientation, notre première contribution a consisté à proposer un modèle de gestion de profils d'interrogation construits à partir des différentes sessions de recherche effectuées par un utilisateur. Nous avons également mené des travaux sur l'évolution des profils à partir des résultats de recherche obtenus afin de maintenir des profils qui correspondent au plus près aux centres d'intérêt de l'utilisateur. Ces travaux ont été menés en lien avec le projet européen IRAIA.

Nous avons ensuite proposé une approche basée sur une gestion de versions dans le but de capitaliser les recherches effectuées par les utilisateurs, de conserver leurs centres d'intérêt et leur évolution. L'intérêt est de conserver une grande finesse dans les informations relatives à l'utilisateur. Ce principe offre des possibilités d'exploitation plus grandes, par exemple en termes d'aide à l'utilisateur dans l'expression de ses besoins. Nous avons ainsi proposé d'offrir à un utilisateur la possibilité d'exprimer son besoin d'information en cheminant dans un graphe construit à partir des requêtes exprimées précédemment par les autres utilisateurs. Cette proposition a été mise en œuvre dans un prototype baptisé *Query Explorer*.

1.4 Quatrième orientation : Adapter le comportement du système

Nous nous sommes intéressés dans cette dernière orientation à la possibilité d'adapter le système en fonction de la recherche effectuée. Une première contribution a été d'étudier comment agir sur les paramètres pour améliorer l'efficacité de notre SRI XML en fonction de préférences de l'utilisateur. Une seconde proposition a consisté en une méthode de combinaison de SRI par sélection du système le plus efficace sur les premiers documents restitués par différents SRI en se

basant sur le retour de pertinence fourni par l'utilisateur. Des expérimentations réalisées sur des collections de référence INEX et TREC ont permis de valider nos propositions.

Dans le cadre de cette quatrième orientation, nous nous sommes également intéressés à l'identification des contextes pour lesquels un système est efficace et ceux pour lesquels il ne l'est pas. Nous avons, d'une part, identifié des critères de distinction des requêtes et leur influence sur l'efficacité de notre SRI XML en s'appuyant sur des expérimentations que nous avons réalisées sur les collections INEX. Nous avons proposé, d'autre part, un cadre pour évaluer l'adéquation d'une interface de restitution de résultat et différents scénarios. Ceci permettra ensuite de définir un système qui adapte la visualisation des résultats en fonction de la recherche effectuée. Un prototype de plateforme d'évaluation baptisé *VSE* implémente cette proposition.

Le projet européen Ws-Talk, le projet Égide Alliance Grande-Bretagne QUEST et le projet ACRIC du programme pluri-formations FRÉMIT ont constitué des cadres applicatifs pour ces travaux.

2 Perspectives

La contextualisation de la recherche d'information n'en est qu'à son commencement et constitue un domaine de recherche à long terme (Allan *et al.*, 2003) réaffirmé au sein de la communauté de la RI (Crestani, 2007; Fuhr, 2010). Le défi est d'offrir à l'utilisateur le système le plus adapté à chaque recherche d'information qu'il effectue. La multiplicité des composants du contexte et leur interdépendance constituent, selon moi, le point central du problème de contextualisation de la RI.

Un premier verrou est lié à la complexité de la notion de contexte elle-même. La multiplicité des aspects du contexte (Cool et Spink, 2002; Ingwersen et Järvelin, 2005; Göker *et al.*, 2009) rend sa modélisation difficile. Les systèmes de RI actuels ne prennent pas en compte l'ensemble de ces aspects ; ils en considèrent au mieux un comme le domaine de l'information (cf. chapitre 1) ou les centres d'intérêts de l'utilisateur (cf. chapitre 3). La maîtrise de la variété des contextes constitue un premier challenge. Il faut être capable de définir des modèles permettant de représenter les différents aspects du contexte en RI. Certains de ces aspects ont été abordés dans le cadre de la RI (Göker *et al.*, 2009), mais aucun modèle ne permet d'intégrer l'ensemble de ces aspects. De nombreux travaux restent à faire pour que la communauté arrive à proposer une nomenclature de ces différents aspects. Les travaux doivent être poursuivis pour modéliser chacun de ces aspects. La modélisation du contexte n'est pas qu'une étape et un deuxième challenge consiste à identifier le contexte lorsque l'utilisateur interagit avec un SRI. Dans la mesure où le contexte peut être représenté, il faut être capable de reconnaître un contexte lorsqu'il se présente pour répondre de façon adaptée au besoin.

Le second verrou est lié à la corrélation entre les éléments du contexte et les composants du processus de RI. En effet, une fois le contexte identifié, le système doit pouvoir sélectionner judicieusement les technologies les plus adéquates par rapport à ce contexte. En RI, un certain nombre de composants ont été définis tels que le modèle de représentation des documents (indexation), le principe de reformulation de requête, le modèle de recherche ou de mise en correspondance, le modèle de restitution. Pour chacun de ces composants, différentes approches ont été proposées dans la littérature. Si l'on considère l'indexation automatique par exemple, certains systèmes se

basent sur l'extraction de termes à partir d'un vocabulaire contrôlé, sur des mécanismes de traitement du langage naturel ou sur la pondération des unités extraites. La même variété se retrouve dans les méthodes de reformulation de requêtes (utilisation de ressources externes ou des documents pertinents...) ou dans les modèles de recherche (vectoriel, probabiliste, sémantique...). Cependant, les contextes dans lesquels chaque composant est intéressant ne sont pas clairement identifiés. Il est donc nécessaire d'évaluer au niveau de quel(s) composant(s) agir et comment étendre ces composants pour prendre en compte les éléments de contexte.

Un troisième verrou est lié à l'interdépendance entre les éléments de contexte. En effet, il est nécessaire de concevoir des solutions complémentaires qui prennent en compte les différents éléments du contexte en évitant d'étendre un SRI avec des solutions incompatibles. Ce n'est qu'alors qu'il sera réellement possible de définir un système adapté au contexte identifié. Comme premier challenge, il faut analyser comment les éléments du contexte et leur coexistence influencent la perception du résultat de recherche par l'utilisateur. Le second challenge est de définir un cadre d'évaluation pour mesurer l'efficacité des propositions de SRI s'adaptant au contexte.

Dans cette optique, je projette mes activités de recherche dans les années futures de la manière suivante :

- à court terme, le projet CAAS (« Contextual Analysis and Adaptive Search ») va offrir un cadre applicatif pour évaluer le degré de généricité de nos contributions en matière de caractérisation des contextes et des SRI. Le fort investissement prévu dans ce projet et plus particulièrement concernant le lot relatif à l'analyse de la variété des systèmes, que je dois prendre en charge, va contribuer à la poursuite de ces travaux. L'objectif est d'analyser les systèmes afin d'identifier les caractéristiques qui distinguent les systèmes ainsi que leurs relations avec les composants du processus de RI. Ces travaux constituent l'étape suivante à nos contributions en s'appuyant notamment sur le prototype de SRI modulaire que nous avons développé, en changeant par exemple de méthode d'indexation, de pondération des termes, de fonction d'appariement. Des SRI libres tels que Terrier [15] seront également étudiés dans ce cadre.

- à moyen terme, étudier une méthode de combinaison de SRI qui s'adapte à chaque requête constitue également une perspective pour répondre au second verrou. Les combinaisons de SRI de la littérature s'appuient sur des caractéristiques des résultats restitués en appliquant un traitement identique pour chaque requête. Notre hypothèse est qu'une amélioration de l'efficacité des SRI passe par la recherche de la manière de combiner judicieusement les SRI à chaque requête en fonction de ses caractéristiques. Pour parvenir à valider cette hypothèse, la caractérisation des requêtes est à approfondir afin de dégager une typologie suivant des caractéristiques formelles. La définition d'une correspondance entre ces caractéristiques formelles des requêtes et les propriétés formelles caractérisant les SRI identifiées dans le cadre du projet CAAS conduira à une nouvelle approche pour combiner les SRI. L'idée est de s'appuyer sur les variantes de notre prototype de SRI définies lors du projet et de définir une méthode qui adapte la combinaison des variantes en fonction de la requête. Une approche complémentaire à explorer est de proposer un principe de type « Learning to rank ». Nos contributions autour d'un modèle de RI adaptable mis en œuvre dans notre prototype de SRI, nous amènent à réfléchir à la définition d'un processus

15. `http://terrier.org/`

d'apprentissage automatique pour identifier la variante la plus adaptée par exemple à une collection donnée.

Pour aller plus loin, je vais poursuivre mes travaux sur les interfaces de restitution de résultats (VIRI) pour proposer des solutions d'adaptation de l'interface en fonction du scénario de recherche. Le prototype de plateforme d'évaluation de VIRI que nous avons développé nous permettra de mettre en place une série de campagnes d'évaluation de VIRI pour identifier les corrélations entre constituants de l'interface et scénarios de recherche et contribuera à répondre au second verrou. Il s'agit également de répondre au premier verrou en proposant une approche pour identifier le scénario correspondant à une recherche en cours. Les aspects cognitifs et ergonomiques de ces travaux permettront de renforcer nos collaborations avec des équipes de recherche en sciences cognitives. La mise en œuvre dans un prototype de système adaptatif au niveau de l'interface de restitution développé au-dessus de notre prototype de SRI et des expérimentations finaliseront ces travaux.

Enfin, je prévois d'aborder le troisième verrou au travers de la recherche multi-dimensions c'est-à-dire la recherche combinant plusieurs dimensions. La recherche multi-dimensions est un cas de RI dans lequel un contexte est constitué d'éléments interdépendants. La recherche géographique est un exemple sur trois dimensions : spatiale, temporelle et thématique. Nous avons initié, en collaboration avec l'équipe T2I [16] de l'université de Pau et des Pays de l'Adour, des travaux proposant un SRI prenant en compte les éléments de contexte interdépendants relatifs à l'information géographique dans le processus de RI et un cadre d'évaluation pour de tels SRI (Palacio *et al.*, 2010b,a; Cabanac *et al.*, 2011b). Ces travaux ouvrent des perspectives de recherche en matière d'adaptation de SRI en fonction d'un contexte impliquant des éléments interdépendants au travers des différentes combinaisons des dimensions dans l'expression du besoin et dans les documents.

Pour mener ces travaux, l'accès à des collections d'information et des méthodes d'évaluation appropriées est indispensable. Dans cette optique, nous avons défini un projet de centre de données interdisciplinaire, au sein duquel je serai responsable du projet pilote RI. L'objectif est de disposer d'une infrastructure qui sous-tende des projets innovants notamment en lien avec la contextualisation de la RI. Il s'agit de constituer des collections de documents et de définir des cadres d'évaluation appropriés (protocoles et mesures) qui manquent actuellement. Ce projet s'inscrit également dans une démarche à long terme d'accès à des ressources pour la RI et l'exploration de données de manière plus générale.

– à long terme, la question de l'interdépendance des composants du contexte constituera un axe de travail. En effet, la difficulté d'adapter le SRI au contexte réside dans l'interdépendance des composants du contexte, leur priorisation voire leur concurrence. Concevoir des solutions d'adaptation en fonction d'éléments du contexte considérés de manière indépendante est insuffisant. La solution d'adaptation du SRI en fonction du contexte doit être envisagée de manière globale. Au-delà, la prise en compte de la dynamique représente une problématique majeure. La dynamique des sources d'information et des utilisateurs sont des thématiques d'avenir selon Dumais (2009). Nous avons initiés des travaux relatif à la dynamique des utilisateurs (au travers de la capitalisation de leurs expériences de

16. http://liuppa.univ-pau.fr/live/EquipesdeRecherche/Equipe_T2I

recherche) ainsi qu'à la dynamique des sources d'information (au travers de la dynamique de l'indexation) et qui doivent être développés et renforcés. Ces aspects s'inscrivent dans le champ plus vaste de la dynamique des contextes qui ajoute à la complexité des contextes la complexité de leur évolution en termes de composants exploitables et d'interdépendance entre ces composants. Les composants d'un contexte disponibles ne sont pas toujours les mêmes et leur interdépendance varie également, il faut donc définir une solution qui prenne en compte cette dynamique. Dans ce cadre, la recherche d'information ambiante représente un domaine d'application particulièrement intéressant car porteur de problématiques liées à la prise en compte de la dynamique au niveau des sources d'information, au niveau des utilisateurs et plus globalement au niveau des contextes.

Pour mener à bien ces activités, je compte mettre à profit l'expérience acquise en matière d'organisation de la recherche c'est-à-dire :

- renforcer et compléter les collaborations existantes avec des partenaires industriels et académiques,
- construire des projets qui valideront les solutions proposées par rapport aux problématiques analysées dans des cadres applicatifs concrets,
- concevoir des prototypes,
- définir les expérimentations visant à mesurer la qualité des résultats obtenus en définissant des cadres d'évaluation et en concevant de nouvelles campagnes d'évaluation.

Bibliographie

A

Airio, E., Järvelin, K., Saatsi, P., Kekäläinen, J. et Suomela, S. (2004). CIRI - An Ontology-based Query Interface for Text Retrieval. *In Proceedings of The Web Intelligence symposium.* Cité 2 fois, p. 10 et 11.

Al Sabbagh, R. (2005). Partage d'expérience de recherche d'information basée sur l'évolution de profils. Mémoire de Master Recherche, IRIT, Université Paul Sabatier, France. Cité 4 fois, p. 72, 85, 160 et 163.

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J. et Zhai, C. (2003). Challenges in information retrieval and language modeling : report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*, 37(1):31–47. Cité 2 fois, p. 2 et 129.

Amardeilh, F. (2006). Ontopop or how to annotate documents and populate ontologies from texts. *In Proceedings of the Workshop on Mastering the Gap : From Information Extraction to Semantic Representation (ESWC'06).* Cité 1 fois, p. 11.

Amato, G. et Straccia, U. (1999). User profile modeling and applications to digital libraries. *In ECDL '99 : Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, pages 184–197, London, UK. Springer-Verlag. Cité 1 fois, p. 65.

Amer-Yahia, S., Lakshmanan, L. V. S. et Pandit, S. (2004). FleXPath : flexible structure and full-text querying for XML. *In SIGMOD '04 : Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 83–94, New York, NY, USA. ACM. Cité 1 fois, p. 37.

Amitay, E., Darlow, A., Konopnicki, D. et Weiss, U. (2005). Queries as anchors : selection by association. *In HYPERTEXT '05 : Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 193–201, New York, NY, USA. ACM. Cité 2 fois, p. 66 et 72.

Andonoff, E., Hubert, G. et Le Parc, A. (1998). A database interface integrating a querying language for versions. *In* Litwin, W., Morzy, T. et Vossen, G., éditeurs : *Advances in Databases and Information Systems, Second East European Symposium, ADBIS'98*, volume 1475 de *Lecture Notes in Computer Science*, pages 200–211. Springer. Cité 1 fois, p. 76.

Andreasen, T., Bulskov, H. et Knappe, R. (2003). Similarity for conceptual querying. *In* Yazici, A. et Sener, C., éditeurs : *Computer and Information Sciences - ISCIS 2003, 18th International Symposium*, volume 2869 de *Lecture Notes in Computer Science*, pages 268–275. Springer. Cité 1 fois, p. 29.

Arezki, R., Poncelet, P., Dray, G. et Pearson, D. W. (2004). Web information retrieval based on user profile. *In* Bra, P. D. et Nejdl, W., éditeurs : *Adaptive Hypermedia and Adaptive Web-Based Systems, Third International Conference, AH 2004*, volume 3137 de *Lecture Notes in Computer Science*, pages 275–278. Springer. Cité 2 fois, p. 65 et 66.

Armstrong, R., Freitag, D., Joachims, T. et Mitchell, T. (1995). WebWatcher : A learning apprentice for the World Wide Web. *In AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pages 6–12, Stanford. Cité 1 fois, p. 65.

Arvola, P., Junkkari, M. et Kekäläinen, J. (2005). Generalized contextualization method for xml information retrieval. *In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 20–27, New York, NY, USA. ACM. Cité 1 fois, p. 37.

Ashoori, E. et Lalmas, M. (2007). Using topic shifts for focussed access to XML repositories. *In ECIR'07 : Proceedings of the 29th European conference on IR research*, pages 444–455, Berlin, Heidelberg. Springer-Verlag. Cité 2 fois, p. 36 et 37.

Aufaure, M.-A., Soussi, R. et Zghal, H. B. (2007). SIRO : On-line semantic information retrieval using ontologies. *In ICDIM '07 : Second IEEE International Conference on Digital Information Management*, pages 321–326. IEEE. Cité 2 fois, p. 8 et 11.

Augé, J., Englmeier, K., Hubert, G. et Josiane, M. (2001). Classification automatique de textes basée sur des hiérarchies de concepts. *In VSST'2001 : Veille Stratégique Scientifique et Technologique*, pages 291–300. Cité 9 fois, p. 12, 15, 17, 18, 20, 33, 160, 162 et 165.

Augé, J., Englmeier, K., Hubert, G. et Mothe, J. (2003). Catégorisation automatique de textes basée sur des hiérarchies de concepts. *In* Chrisment, C., éditeur : *BDA '03 : 19èmes Journées Bases de Données Avancées.* Cité 8 fois, p. 12, 15, 21, 22, 33, 160, 162 et 165.

Aussenac-Gilles, N. et Mothe, J. (2004). Ontologies as background knowledge to explore document collections. *In* Fluhr, C., Grefenstette, G. et Croft, W. B., éditeurs : *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2004, 7th International Conference*, pages 129–142. CID. Cité 3 fois, p. 7, 23 et 29.

B

Baccini, A., Déjean, S., Kompaoré, D. et Mothe, J. (2010). Analyse des critères d'évaluation des systèmes de recherche d'information. *Technique et Science Informatiques*, 29(3):289–308. Cité 1 fois, p. 93.

Bachimont, B. (2000). *Engagement Sémantique et Engagement Ontologique : Conception et Réalisation D'ontologies En Ingénierie Des Connaissances*, chapitre 19, pages 305–324. Cité 1 fois, p. 29.

Baeza-Yates, R. A., Hurtado, C. A. et Mendoza, M. (2004). Query recommendation using query logs in search engines. *In* Lindner, W., Mesiti, M., Türker, C., Tzitzikas, Y. et Vakali, A., éditeurs : *EDBT Workshops*, volume 3268 de *Lecture Notes in Computer Science*, pages 588–596. Springer. Cité 2 fois, p. 64 et 66.

Baeza-Yates, R. A. et Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. Cité 1 fois, p. 69.

Balfe, E. et Smyth, B. (2005). An analysis of query similarity in collaborative Web search. *In* Losada, D. E. et Fernández-Luna, J. M., éditeurs : *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005*, volume 3408 de *Lecture Notes in Computer Science*, pages 330–344. Springer. Cité 2 fois, p. 65 et 66.

Banks, D., Over, P. et Zhang, N.-F. (1999). Blind men and elephants : Six approaches to TREC data. *Inf. Retr.*, 1(1-2):7–34. Cité 2 fois, p. 88 et 92.

Barde, J., Divol, J., Libourel, T. et Maurel, P. (2005a). Interface adaptable de requêtes pour un service de métadonnées. *Revue des Nouvelles Technologies de l'Information*, RNTI E5 - Extraction des connaissances : État et perspective:135–140. Cité 1 fois, p. 9.

Barde, J., Libourel, T. et Maurel, P. (2005b). A metadata service for integrated management of knowledges related to coastal areas. *Multimedia Tools and Applications*, 25(3):419–429. Cité 1 fois, p. 9.

Beitzel, S. M., Frieder, O., Jensen, E. C., Grossman, D., Chowdhury, A. et Goharian, N. (2003). Disproving the fusion hypothesis : an analysis of data fusion via effective information retrieval strategies. *In SAC '03 : Proceedings of the 2003 ACM symposium on Applied computing*, pages 823–827, New York, NY, USA. ACM. Cité 1 fois, p. 90.

Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., jeng Lin, S., Lobash, L., Park, S., Savage-Knepshield, P. A. et Sikora, C. (1999). Relevance feedback *versus* local context analysis as term suggestion devices : Rutgers' TREC-8 Interactive Track experience. *In TREC-8 : The 8th Text REtrieval Conference*. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 71.

Belkin, N. J., Cool, C., Koenemann, J., Ng, K. B. et Park, S. (1996). Using relevance feedback and ranking in interactive searching. *In TREC-4 : The 4th Text Retrieval Conference*, pages 181–209. National Institute of Standards and Technology (NIST). Cité 2 fois, p. 64 et 116.

Benammar, A. (2003). *Profils en recherche d'information : définition, exploitation et adaptation*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France. Cité 4 fois, p. 67, 85, 160 et 163.

Benammar, A., Boughanem, M., Hubert, G., Laffaire, C. et Mothe, J. (2002a). IRIT at TREC'2002 :Web Track. *In TREC-11 : The 11th Text REtrieval Conference, Gaithersburg*. National Institute of Standards and Technology (NIST). Cité 4 fois, p. 61, 160, 161 et 163.

Benammar, A., Hubert, G. et Mothe, J. (2001). Reformulation automatique des profils utilisant un ensemble local de documents. *In VSST'2001 : Veille Stratégique Scientifique et Technologique*, pages 321–329. Cité 5 fois, p. 67, 85, 160, 162 et 163.

Benammar, A., Hubert, G. et Mothe, J. (2002b). Automatic profile reformulation using a local document analysis. *In* Crestani, F., Girolami, M. et van Rijsbergen, C. J., éditeurs : *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research*, volume 2291 de *Lecture Notes in Computer Science*, pages 124–134. Springer. Cité 7 fois, p. 67, 69, 85, 160, 161, 163 et 165.

Benammar, A., Hubert, G. et Mothe, J. (2003). Proposition à l'intégration des profils dans le processus de recherche d'information. *ISDM '03 : International Journal of Information Sciences for Decision Making*, 6(47):143–152. Cité 6 fois, p. 67, 68, 85, 160, 161 et 163.

Berisha-Bohé, S. et Rumpler, B. (2007). Modèle évolutif d'un profil utilisateur. *In COnférence en Recherche d'Infomations et Applications - CORIA 2007, 4th French Information Retrieval Conference*, pages 197–210. Cité 1 fois, p. 65.

Bhogal, J., Macfarlane, A. et Smith, P. (2007). A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886. Cité 2 fois, p. 8 et 11.

Bodner, R. C. et Song, F. (1996). Knowledge-based approaches to query expansion in information retrieval. *In AI '96 : Proceedings of the 11th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, pages 146–158, London, UK. Springer-Verlag. Cité 1 fois, p. 8.

Bonnel, N., Chevalier, M., Dousset, B. et Hubert, G. (2008). Visualisation en recherche d'information. *In* Chauvin, S., éditeur : *Information & visualisation – enjeux, recherches et applications*. Cépaduès Éditions. Cité 4 fois, p. 113, 114, 125 et 162.

Bonnel, N., Lemaire, V., Cotarmanac'h, A. et Morin, A. (2006). Effective organization and visualization of web search results. *In IMSA'06 : Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pages 209–216, Anaheim, CA, USA. ACTA Press. Cité 2 fois, p. 91 et 92.

Bouchard, H. et Nie, J.-Y. (2006). Modèles de langue appliqués à la recherche d'information contextuelle. *In COnférence en Recherche d'Infomations et Applications - CORIA 2006, 3rd French Information Retrieval Conference*, pages 213–224. Cité 1 fois, p. 8.

Bray, T., Paoli, J. et Sperberg-McQueen, C. M. (1998). Extensible Markup Language (XML) 1.0 W3C Recommendation. Cité 1 fois, p. 35.

Brewster, C. et Wilks, Y. (2004). *In* Deegan, M., éditeur : *The Keyword Project : Unlocking Content through Computational Linguistics*. Centre for Computing in the Humanities, Kings College London. Cité 1 fois, p. 29.

Brown, C. M. (1988). *Human-computer interface design guidelines*. Ablex Publishing Corp., Norwood, NJ, USA. Cité 1 fois, p. 92.

Büttcher, S., Clarke, C. L. A. et Cormack, G. V. (2004). Domain-specific synonym expansion and validation for biomedical information retrieval. *In TREC-13 : the 13th Text REtrieval Conference*. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 8.

Buckley, C., Salton, G. et Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. *In SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 292–300, New York, NY, USA. Springer-Verlag New York, Inc. Cité 1 fois, p. 69.

Buxton, S. et Rys, M. (2003). XQuery and XPath full-text requirements. Cité 1 fois, p. 44.

C

Cabanac, G. (2008). *Fédération et amélioration des activités documentaires par la pratique d'annotation collective*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France. Cité 1 fois, p. 66.

Cabanac, G., Hubert, G., Boughanem, M. et Chrisment, C. (2010a). Impact du « biais des *ex aequo* » dans les évaluations de Recherche d'Information. *In CORIA'10 : Actes de la 7e conférence en recherche d'information et applications*, pages 83–98. Cité 2 fois, p. 61 et 161.

Cabanac, G., Hubert, G., Boughanem, M. et Chrisment, C. (2010b). Tie-breaking Bias : Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. *In* Agosti, M., Ferro, N., Peters, C., de Rijke, M. et Smeaton, A. F., éditeurs : *CLEF'10 : Proceedings of the 1st Conference on Multilingual and Multimodal Information Access Evaluation*, volume 6360 de *Lecture Notes In Computer Science*, pages 112–123. Springer-Verlag. Cité 2 fois, p. 61 et 161.

Cabanac, G., Hubert, G., Boughanem, M. et Chrisment, C. (2011a). Impact du « biais des ex aequo » dans les évaluations de Recherche d'Information. *Document numérique*, 14:(à paraître). version étendue de CORIA'10. Cité 2 fois, p. 61 et 161.

Cabanac, G., Palacio, D., Sallaberry, C. et Hubert, G. (2011b). Évaluation de la pertinence des résultats en recherche d'information géographique : définition d'un cadre expérimental et validation de l'apport des dimensions de l'information géographique. *Document numérique*, 14:(à paraître). version étendue de INFORSID'10. Cité 5 fois, p. 33, 34, 61, 131 et 161.

Calegari, S. et Pasi, G. (2008). Personalized ontology-based query expansion. *In WI-IAT '08 : Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 256–259, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 66.

Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y. et Hon, H.-W. (2006). Adapting ranking SVM to document retrieval. *In SIGIR '06 : Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193, New York, NY, USA. ACM. Cité 2 fois, p. 89 et 93.

Carman, M. J., Baillie, M. et Crestani, F. (2008). Tag data and personalized information retrieval. *In SSM '08 : Proceeding of the 2008 ACM workshop on Search in social media*, pages 27–34, New York, NY, USA. ACM. Cité 2 fois, p. 64 et 66.

Carmel, D., Maarek, Y. S., Mandelbrod, M., Mass, Y. et Soffer, A. (2003). Searching XML documents via XML fragments. *In SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 151–158, New York, NY, USA. ACM. Cité 3 fois, p. 37, 48 et 50.

Carmel, D., Yom-Tov, E., Darlow, A. et Pelleg, D. (2006). What makes a query difficult ? *In SIGIR '06 : Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 390–397, New York, NY, USA. ACM. Cité 1 fois, p. 88.

Chang, Y.-C., Chen, S.-M. et Liau, C.-J. (2008). Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method. *Expert Syst. Appl.*, 34(3):1948–1953. Cité 1 fois, p. 11.

Chen, C. (2006). *Information Visualization : Beyond the Horizon*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. Cité 1 fois, p. 90.

Chen, H. et Zeng, C. (2008). Personalized information retrieval model based on user interests. *In CSSE '08 : Proceedings of the 2008 International Conference on Computer Science and Software Engineering*, pages 1063–1066, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 66.

Chevalier, M. (2002). *Interface adaptative pour l'aide à la recherche d'information sur le web*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France. Cité 2 fois, p. 91 et 92.

Chevalier, M. et Hubert, G. (2005). Évaluation d'une interface de restitution de recherche : Quelles conclusions en tirer ? *In Atelier Visualisation et Extraction de Connaissances - EGC'05*, pages 15–25. Cité 4 fois, p. 113, 114, 125 et 162.

Chevalier, M. et Hubert, G. (2009). Query assistant based on experience capitalization for information retrieval systems. *In HSI'09 : Proceedings of the 2nd conference on Human System Interactions*, pages 499–506, Piscataway, NJ, USA. IEEE Press. Cité 8 fois, p. 78, 81, 82, 83, 84, 85, 161 et 165.

Chevalier, M., Soulé-Dupuy, C. et Tchienehom, P. (2005). Profiles re-usability for personalized information access : application to users contexts determination. *In CIR'05 : Proceedings of the International Workshop on Context-based Information Retrieval, in conjunction with Context'05*, pages 71–82. Cité 1 fois, p. 65.

Chin, K. K. (1999). Support vector machines applied to speech pattern classification. Master's thesis, Cambridge University Engineering Department, Cambridge, UK. Cité 1 fois, p. 10.

Chrisment, C., Hernandez, N., Genova, F. et Mothe, J. (2006a). D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus. *AMETIST*, 0:59–92. Cité 1 fois, p. 30.

Chrisment, C., Hernandez, N., Hubert, G. et Mothe, J. (2006b). Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents. *Information – Interaction – Intelligence*, Numéro spécial Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance:53–83. Cité 3 fois, p. 30, 33 et 161.

Clark, J. et DeRose, S. (1999). XML Path Language (XPath) Version 1.0. Cité 3 fois, p. 41, 42 et 52.

Clarke, C. L. A. (2005). Controlling overlap in content-oriented XML retrieval. *In SIGIR '05 : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 314–321, New York, NY, USA. ACM. Cité 1 fois, p. 38.

Clarke, C. L. A. et Tilker, P. L. (2005). Multitext experiments for inex 2004. *In Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 85–87. Springer. Cité 1 fois, p. 110.

Cleverdon, C. W. (1967). The Cranfield tests on English language devices. *Aslib Proceedings*, 19(6):173–194. Cité 2 fois, p. 1 et 88.

Cohen, A. M., Bhupatiraju, R. T. et Hersh, W. R. (2004). Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. *In TREC-13 : The 13th Text REtrieval Conference*. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 89.

Conradi, R. et Westfechtel, B. (1998). Version models for software configuration management. *ACM Comput. Surv.*, 30(2):232–282. Cité 1 fois, p. 76.

Cool, C. et Spink, A. (2002). Issues of context in information retrieval (ir) : an introduction to the special issue. *Inf. Process. Manage.*, 38(5):605–611. Cité 1 fois, p. 129.

Crestani, F. (2007). Information retrieval in context. Talk at DBTA Workshop on Mobile Information : Context-Awareness. Cité 1 fois, p. 129.

Cronen-Townsend, S., Zhou, Y. et Croft, W. B. (2002). Predicting query performance. *In SIGIR '02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, New York, NY, USA. ACM. Cité 1 fois, p. 88.

Crouch, C. J., Khanna, S., Potnis, P. et Doddapaneni, N. (2006). The dynamic retrieval of XML elements. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 268–281, New York, NY, USA. Springer-Verlag. Cité 1 fois, p. 36.

Cugini, J. V., Laskowski, S. et Sebrechts, M. M. (2000). Design of 3D visualization of search results : evolution and evaluation. volume 3960, pages 198–210. SPIE. Cité 1 fois, p. 91.

Cui, H., Wen, J.-R. et Chua, T.-S. (2003a). Hierarchical indexing and flexible element retrieval for structured document. *In ECIR'03 : Proceedings of the 25th European conference on IR research*, pages 73–87, Berlin, Heidelberg. Springer-Verlag. Cité 1 fois, p. 45.

Cui, H., Wen, J.-R., Nie, J.-Y. et Ma, W.-Y. (2003b). Query expansion by mining user logs. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):829–839. Cité 2 fois, p. 64 et 66.

D

Danilowicz, C. et Nguyen, H. C. (2002). Using user profiles in intelligent information retrieval. *In ISMIS '02 : Proceedings of the 13th International Symposium on Foundations of Intelligent Systems*, pages 223–231, London, UK. Springer-Verlag. Cité 2 fois, p. 65 et 66.

Dasgupta, A., Drineas, P., Harb, B., Josifovski, V. et Mahoney, M. W. (2007). Feature selection methods for text classification. *In KDD '07 : Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 230–239, New York, NY, USA. ACM. Cité 1 fois, p. 89.

Dash, M. et Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156. Cité 1 fois, p. 89.

de Vries, A. P., Kazai, G. et Lalmas, M. (2004). Evaluation metrics 2004. *In* Fuhr, N., Lalmas, M., Malik, S. et Szlávik, Z., éditeurs : *Pre Proceedings of the 3rd INEX Workshop*, pages 249–250. Cité 1 fois, p. 52.

Denos, N. (1997). *Modélisation de la pertinence en recherche d'information : modèle conceptuel, formalisation et application*. Thèse de doctorat, Université Joseph-Fourier - Grenoble I, Grenoble, France. Cité 1 fois, p. 1.

Denos, N., Berrut, C., Gallardo-López, L. et Nguyen, A.-T. (2004). Cocofil : une plateforme de filtrage collaboratif orientée vers la communauté. *In CORIA '04 : COnférence en Recherche d'Information et Application*, pages 9–26. Cité 1 fois, p. 66.

Desmontils, E. et Jacquin, C. (2002). Indexing a Web site with a terminology oriented ontology. *In The Emerging Semantic Web*, pages 181–198. IOS Press. Cité 2 fois, p. 11 et 28.

Díaz-Galiano, M. C., Martín-Valdivia, M. et Ureña-López, L. A. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. *Comput. Biol. Med.*, 39(4):396–403. Cité 1 fois, p. 8.

Dimililer, N., Varoğlu, E. et Altinçay, H. (2007). Vote-based classifier selection for biomedical NER using genetic algorithms. *In IbPRIA '07 : Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II*, pages 202–209, Berlin, Heidelberg. Springer-Verlag. Cité 1 fois, p. 89.

Dkaki, T., Hubert, G., Mothe, J. et Orain, E. (2004). Recherche de la nouveauté dans les textes : une tâche difficile. *In VSST'2004 : Veille Stratégique Scientifique et Technologique*, volume 2, pages 355–368. Cité 3 fois, p. 85, 100 et 162.

Dudognon, D., Hernandez, N., Hubert, G., Lardy, C., Laublet, P., Mothe, J., Ralalason, B., Tissaoui, A., Reymonet, A. et Thomas, J. (2010a). Livrable 5.2 : Spécifications des outils d'annotation dynamique et de recherche dans les textes. Rapport technique Projet ANR Dynamo 07 TLOG 004 01, ANR. Cité 2 fois, p. 156 et 162.

Dudognon, D., Hubert, G. et Ralalason, B. (2010b). PROXIGÉNÉA : Une mesure de similarité conceptuelle. *In VSST '10 : Colloque Veille Stratégique Scientifique et Technologique*, page (support électronique). Cité 4 fois, p. 33, 156, 161 et 163.

Duke, A., Glover, T. et Davies, J. (2007). Squirrel : An advanced semantic search and browse facility. *In ESWC '07 : Proceedings of the 4th European conference on The Semantic Web*, pages 341–355, Berlin, Heidelberg. Springer-Verlag. Cité 1 fois, p. 9.

Dumais, S. et Chen, H. (2000). Hierarchical classification of Web content. *In SIGIR '00 : Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263, New York, NY, USA. ACM. Cité 1 fois, p. 10.

Dumais, S. T. (2009). An interdisciplinary perspective on information retrieval. *In SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 1–2, New York, NY, USA. ACM. Cité 1 fois, p. 131.

Dumas, J. S. et Redish, J. C. (1999). *A Practical Guide to Usability Testing*. Intellect Books, Exeter, UK, UK. Cité 1 fois, p. 92.

E

Eguchi, K. (2000). Incremental query expansion using local information of clusters. *In SCI 2000 : Proceedings of the 4th World Multiconference on Systemics, Cybernetics and Informatics*, volume 2, pages 310–316. Cité 1 fois, p. 69.

El Makssoud, M. (2010). Partage et capitalisation d'expérience pour la recherche d'information. Mémoire de Master Recherche, IRIT, Université Paul Sabatier, France. Cité 3 fois, p. 85, 160 et 163.

Elsas, J. L., Carvalho, V. R. et Carbonell, J. G. (2008). Fast learning of document ranking functions with the committee perceptron. *In WSDM '08 : Proceedings of the international conference on Web search and web data mining*, pages 55–64, New York, NY, USA. ACM. Cité 2 fois, p. 89 et 93.

Englmeier, K., Hubert, G. et Mothe, J. (2006). Distinguer les requêtes pour améliorer la recherche d'information XML. *In COnférence en Recherche d'Infomations et Applications - CORIA 2006, 3rd French Information Retrieval Conference*, pages 41–52. Cité 5 fois, p. 106, 107, 125, 157 et 162.

Englmeier, K. et Mothe, J. (2002). Final report - estage. Rapport technique Rapport final de contrat, eStage, EC. Cité 1 fois, p. 12.

F

Fan, W., Gordon, M. D. et Pathak, P. (2004a). Discovery of context-specific ranking functions for effective information retrieval using genetic programming. *IEEE Trans. on Knowl. and Data Eng.*, 16(4):523–527. Cité 1 fois, p. 89.

Fan, W., Luo, M., Wang, L., Xi, W. et Fox, E. A. (2004b). Tuning before feedback : combining ranking discovery and blind feedback for robust retrieval. *In SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–145, New York, NY, USA. ACM. Cité 1 fois, p. 89.

Fekete, J.-D. et Plaisant, C. (2004). Les leçons tirées des deux compétitions de visualisation d'information. *In IHM 2004 : Proceedings of the 16th conference on Association Francophone d'Interaction Homme-Machine*, pages 7–12, New York, NY, USA. ACM. Cité 1 fois, p. 92.

Fidel, R. (1991). Searchers' selection of search keys : Ii. controlled vocabulary or free-text searching. *JASIS*, 42(7):501–514. Cité 1 fois, p. 22.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. et Ruppin, E. (2001). Placing search in context : the concept revisited. *In WWW '01 : Proceedings of the 10th international conference on World Wide Web*, pages 406–414, New York, NY, USA. ACM. Cité 1 fois, p. 2.

Fitzpatrick, L. et Dent, M. (1997). Automatic feedback using past queries : social searching ? *In SIGIR '97 : Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 306–313, New York, NY, USA. ACM. Cité 2 fois, p. 66 et 77.

Fox, E. A. et Shaw, J. A. (1994). Combination of multiple searches. *In The Second Text REtrieval Conference (TREC-2)*, volume 500-215 de *NIST Special Publication*, pages 243–252. NIST. Cité 7 fois, p. 22, 26, 27, 28, 72, 90 et 100.

Freund, Y., Iyer, R., Schapire, R. E. et Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969. Cité 1 fois, p. 90.

Friberg, K. (2007). Query expansion using domain information in compounds. *In Proceedings of the NAACL-HLT 2007 Doctoral Consortium*, pages 1–4, Rochester, New York. Association for Computational Linguistics. Cité 1 fois, p. 8.

Fu, L., Goh, D. H.-L., Foo, S. S.-B. et Supangat, Y. (2004). Collaborative querying for enhanced information retrieval. *In* Heery, R. et Lyon, L., éditeurs : *Research and Advanced Technology for Digital Libraries, 8th European Conference, ECDL 2004*, volume 3232 de *Lecture Notes in Computer Science*, pages 378–388. Springer. Cité 1 fois, p. 8.

Fugmann, R. (2002). The complementarity of natural and index language in the field of information supply : an overview of their specific capabilities and limitations. *Knowledge Organization*, 29(3-4):217–230. Cité 1 fois, p. 22.

Fuhr, N. (2010). IR Between Science and Engineering, and the Role of Experimentation. Keynote talk at CLEF 2010, Padua, Italy. Cité 1 fois, p. 129.

Fuselier, J. et Chidlovskii, B. (2006). Traitements automatiques pour la migration de documents numériques vers XML. 9(1):9–24. Cité 2 fois, p. 3 et 35.

G

Gantz, J. et Reinsel, D. (2009). As the economy contracts, the Digital Universe expands. Rapport technique IDC Multimedia White Paper. Cité 1 fois, p. 1.

Gery, M., Largeron, C. et Thollard, F. (2008). Integrating structure in the probabilistic model for information retrieval. *In WI-IAT '08 : Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 763–769, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 36.

Geva, S. (2006). GPX - Gardens Point XML IR at INEX 2005. *In Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005, Revised Selected Papers*, volume 3977 de *Lecture Notes in Computer Science*, pages 240–253. Springer. Cité 2 fois, p. 36 et 37.

Gils, B. v. et Schabell, E. (2003). User-profiles for information retrieval. *In The 15th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'03)*, pages 139–146. Cité 1 fois, p. 65.

Göker, A. et McCluskey, T. L. (1991). Towards an adaptive information retrieval system. *In ISMIS '91 : Proceedings of the 6th International Symposium on Methodologies for Intelligent Systems*, volume 542 de *Lecture Notes in Computer Science*, pages 348–357. Springer. Cité 2 fois, p. 65 et 66.

Göker, A., Myrhaug, H. et Bierig, R. (2009). *Context and Information Retrieval*, pages 131–157. Wiley & Sons, Ltd. Cité 1 fois, p. 129.

Grabs, T. et Schek, H.-J. (2002). Generating vector spaces on-the-fly for flexible XML retrieval. *In XML and Information Retrieval Workshop, 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–13. Cité 1 fois, p. 37.

Guha, R., McCool, R. et Miller, E. (2003). Semantic search. *In WWW '03 : Proceedings of the 12th international conference on World Wide Web*, pages 700–709, New York, NY, USA. ACM. Cité 2 fois, p. 8 et 29.

Gövert, N., Fuhr, N., Abolhassani, M. et Großjohann, K. (2003). Content-oriented XML retrieval with HyREX. *In Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, pages 26–32. Cité 1 fois, p. 37.

H

Hadjouni, M., Haddad, M. R., Baazaoui, H., Aufaure, M.-A. et Ben Ghezala, H. (2009). Personalized information retrieval approach. *In WISM '09 : Proceedings of the 6th International Workshop on Web Information Systems Modeling, held in conjunction with CAISE'09*. Cité 1 fois, p. 66.

Hamza, M. B. (2007). Mécanismes de bouclage de pertinence dans un système de recherche d'information XML. Mémoire de Master Recherche, IRIT, Université Paul Sabatier, Toulouse, France. Cité 3 fois, p. 61, 158 et 163.

Handschuh, S., Staab, S. et Ciravegna, F. (2002). S-CREAM - Semi-automatic CREAtion of Metadata. *In EKAW '02 : Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 358–372, London, UK. Springer-Verlag. Cité 1 fois, p. 10.

Harman, D. (1992). Overview of the First Text REtrieval Conference (TREC-1). *In TREC-1 : The 1st Text REtrieval Conference*, pages 1–20. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 1.

Harman, D. et Buckley, C. (2004). The NRRC reliable information access (RIA) workshop. *In SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 528–529, New York, NY, USA. ACM. Cité 3 fois, p. 88, 92 et 100.

Hatano, K., Kinutani, H., Amagasa, T., Mori, Y., Yoshikawa, M. et Uemura, S. (2005). Analyzing the properties of XML fragments decomposed from the INEX document collection. *In Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 de *Lecture Notes in Computer Science*, pages 168–182. Springer. Cité 1 fois, p. 37.

Hayashi, Y., Tomita, J. et Kikui, G. (2000). Searching text-rich XML documents with relevance ranking. *In XML and Information Retrieval Workshop, 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Cité 1 fois, p. 41.

He, B. et Ounis, I. (2003a). A study of parameter tuning for term frequency normalization. *In CIKM '03 : Proceedings of the twelfth international conference on Information and knowledge management*, pages 10–16, New York, NY, USA. ACM. Cité 1 fois, p. 89.

He, B. et Ounis, I. (2003b). University of Glasgow at the Robust Track – a query-based model selection approach for the poorly – performing queries. *In TREC*, pages 636–645. Cité 1 fois, p. 88.

He, D. et Göker, A. (2000). Detecting session boundaries from Web user logs. *In ECIR '00 : Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pages 57–66. Cité 1 fois, p. 80.

Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press, 1 édition. Cité 2 fois, p. 90 et 116.

Hearst, M. A. et Karadi, C. (1997). Cat-a-Cone : an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. pages 246–255. Cité 5 fois, p. 9, 10, 11, 23 et 91.

Hernandez, N., Hubert, G., Lardy, C., Laublet, P., Mothe, J. et Ralalason, B. (2009). Livrable 5.1 : Spécifications des outils d'annotation dynamique et de recherche dans les textes. Rapport technique Projet ANR Dynamo 07 TLOG 004 01, ANR. Cité 2 fois, p. 156 et 162.

Hernandez, N., Mothe, J. et Poulain, S. (2005). Customizing information access according to domain and task knowledge : the ontoExplo system. *In SIGIR '05 : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 607–608, New York, NY, USA. ACM. Cité 2 fois, p. 23 et 29.

Hiemstra, D. (2003). Statistical language models for intelligent xml retrieval. *In* Blanken, H., Grabs, T., Schek, H.-J., Schenkel, R. et Weikum, G., éditeurs : *Intelligent Search on XML Data : Applications, Languages, Models, Implementations, and Benchmarks*, volume 2818 de *Lecture Notes in Computer Science*, pages 107–118. Springer, Berlin, Germany. Cité 1 fois, p. 36.

Hölscher, C. et Strube, G. (2000). Web search behavior of Internet experts and newbies. *In Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 337–346, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co. Cité 2 fois, p. 114 et 115.

Houston, B. et Jacobson, Z. (2000). A simple 3D visual text retrieval interface. Cité 1 fois, p. 91.

Hubert, G. (1997). *Les Versions dans les Bases de Données Orientées Objets : Modélisation et Manipulation.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France. Cité 2 fois, p. 34 et 77.

Hubert, G. (2005). A voting method for XML retrieval. *In Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 183–196. Springer. Cité 6 fois, p. 38, 43, 61, 158, 159 et 161.

Hubert, G. (2006). XML retrieval based on direct contribution of query components. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 172–186, New York, NY, USA. Springer-Verlag. Cité 5 fois, p. 38, 43, 61, 159 et 161.

Hubert, G. (2007). Tuning and evolving retrieval engine by training on previous INEX testbeds. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 5th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006)*, volume 4518 de *Lecture Notes in Computer Science*, pages 243–252, New York, NY, USA. Springer-Verlag. Cité 5 fois, p. 38, 43, 61, 158 et 161.

Hubert, G., Kompaore, N. D. et Mothe, J. (2007a). Réinjection de pertinence pour la fusion de systèmes. *In VSST '07 : Colloque Veille Stratégique Scientifique et Technologique*, page (support électronique). Cité 6 fois, p. 100, 105, 125, 157, 162 et 168.

Hubert, G., Loiseau, Y. et Mothe, J. (2007b). Étude de différentes fonctions de fusion de systèmes de recherche d'information. *In CIDE 10 : Le document numérique dans le monde de la science et de la recherche*, pages 199–207, http ://www.europia.org. EUROPIA. Cité 4 fois, p. 22, 125, 157 et 162.

Hubert, G. et Mothe, J. (2006). Identifier des critères distinguant les requêtes pour améliorer la RI-XML. *Information - Interaction - Intelligence*, 6(2):99–123. Cité 5 fois, p. 106, 107, 125, 157 et 161.

Hubert, G. et Mothe, J. (2007a). Relevance feedback as an indicator to select the best search engine. *In ICEIS '07 : International Conference on Enterprise Information Systems*, pages 184–189. INSTICC Press. Cité 9 fois, p. 22, 100, 102, 103, 125, 157, 161, 167 et 168.

Hubert, G. et Mothe, J. (2007b). Reusing past queries to facilitate information retrieval. *In ICSOFT '07 : International Conference on Software and Data Technologies*, volume 3, pages 166–171. INSTICC Press. Cité 9 fois, p. 22, 67, 72, 73, 74, 76, 85, 161 et 165.

Hubert, G. et Mothe, J. (2009). An adaptable search engine for multimodal information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 60(8):1625–1634. Cité 10 fois, p. 12, 15, 22, 24, 25, 27, 28, 33, 161 et 165.

Hubert, G., Mothe, J., Benammar, A., Dkaki, T., Dousset, B. et Karouach, S. (2001). Textual document mining using graphical interface. *In Universal Access In HCI : Towards an Information Society for All, Proceedings of HCI International '2001 (the 9th International Conference on Human-Computer Interaction)*, volume 1, pages 918–922. Lawrence Erlbaum. Cité 4 fois, p. 62, 160, 161 et 163.

Hubert, G., Mothe, J. et Englmeier, K. (2007c). Tuning search engine to fit XML retrieval scenario. *In WEBIST '07 : International Conference on Web Information Systems and Technologies*, pages 228–233. INSTICC Press. Cité 7 fois, p. 97, 98, 99, 125, 157, 161 et 167.

Hubert, G., Mothe, J. et Poulain, S. (2005). Recherche d'information XML utilisant un principe de vote. *In CORIA '05 : COnférence en Recherche d'Information et Application*, pages 151–166. Cité 3 fois, p. 61, 158 et 162.

Hubert, G., Mothe, J., Ralalason, B. et Ramanonjisoa, B. (2009a). Modèle d'indexation dynamique à base d'ontologies. *In CORIA '09 : COnférence en Recherche d'Infomations et Applications, 6th French Information Retrieval Conference*, pages 169–184. Cité 7 fois, p. 30, 31, 33, 156, 161, 163 et 165.

Hubert, G., Mothe, J. et Will, P. (2009b). Produire une ontologie à partir d'un thésaurus : méthode et outil. ***In VSST '09 : Colloque Veille Stratégique Scientifique et Technologique*****, page (support électronique).** Cité 3 fois, p. 30, 33 et 161.

Hubert, G. et Teste, O. (2009). Analyse multigraduelle OLAP. ***In EGC'2009 : Extraction et gestion des connaissances*****, volume RNTI-E-15 de** ***Revue des Nouvelles Technologies de l'Information*****, pages 241–252. Cépaduès-éditions.** Cité 3 fois, p. 61, 62 et 161.

Hubert, G. et Teste, O. (2010). Multigranularity manipulations for OLAP querying. ***In AKDM : Advances in Knowledge Discovery and Management*****, volume 292 de** ***Studies in Computational Intelligence*****, pages 97–112. Springer, Berlin, Germany.** Cité 3 fois, p. 61, 62 et 161.

I

Ingwersen, P. et Järvelin, K. (2005). *The Turn : Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. Cité 1 fois, p. 129.

Iszlai, Z. et Egyed-Zsigmond, E. (2006). User centered image management system for digital libraries. *In DIAL '06 : Proceedings of the Second International Conference on Document Image Analysis for Libraries*, pages 164–171, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 65.

Iwayama, M. (2000). Relevance feedback with a small number of relevance judgements : incremental relevance feedback vs. document clustering. *In SIGIR '00 : Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–16, New York, NY, USA. ACM. Cité 1 fois, p. 64.

J

Jaimes, A. et Chang, S. (2000). Automatic selection of visual features and classifiers. *In* M. M. Yeung, B.-L. Yeo, . C. A. B., éditeur : *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 3972 de *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 346–358. Cité 1 fois, p. 89.

Jeon, H., Kim, T. et Choi, J. (2008). Adaptive user profiling for personalized information retrieval. *In ICCIT '08 : Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology*, pages 836–841, Washington, DC, USA. IEEE Computer Society. Cité 2 fois, p. 65 et 66.

Joachims, T. (1997). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. *In ICML '97 : Proceedings of the Fourteenth International Conference on Machine Learning*, pages 143–151, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Cité 1 fois, p. 10.

Joachims, T. (1998). Text categorization with support vector machines : learning with many relevant features. *In* Nédellec, C. et Rouveirol, C., éditeurs : *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Heidelberg et al. Springer. Cité 3 fois, p. 10, 11 et 21.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *In KDD '02 : Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA. ACM. Cité 2 fois, p. 64 et 90.

Jomier, G. et Cellary, W. (2000). The database version approach. *Networking and Information Systems Journal*, 3(1):177–214. Cité 1 fois, p. 76.

Jones, G. J. F. et Brown, P. J. (2004). The role of context in information retrieval. *In Workshop on Information Retrieval in Context, the 27th Annual International ACM SIGIR Conference*, pages 133–142. Cité 1 fois, p. 1.

Jones, R. et Klinkner, K. L. (2008). Beyond the session timeout : automatic hierarchical segmentation of search topics in query logs. *In CIKM '08 : Proceeding of the 17th ACM conference on Information and knowledge management*, pages 699–708, New York, NY, USA. ACM. Cité 1 fois, p. 80.

Jéribi, L. et Rumpler, B. (2002). Instance cooperative memory to improve query expansion in information retrieval systems. *Journal of Universal Computer Science*, 8(6):91–601. Cité 1 fois, p. 65.

Järvelin, K. et Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446. Cité 1 fois, p. 53.

Julien, C.-A., Leide, J. E. et Bouthillier, F. (2008). Controlled user evaluations of information visualization interfaces for text retrieval : Literature review and meta-analysis. *J. Am. Soc. Inf. Sci. Technol.*, 59(6):1012–1024. Cité 2 fois, p. 93 et 113.

Jung, S., Herlocker, J. L. et Webster, J. (2007). Click data as implicit relevance feedback in web search. *Inf. Process. Manage.*, 43(3):791–807. Cité 1 fois, p. 64.

K

Kamps, J., de Rijke, M. et Sigurbjörnsson, B. (2004). Length normalization in XML retrieval. *In SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 80–87, New York, NY, USA. ACM. Cité 2 fois, p. 36 et 37.

Kamps, J., Koolen, M. et Sigurbjörnsson, B. (2007). Filtering and clustering XML retrieval results. *In Comparative Evaluation of XML Information Retrieval Systems*, volume 4518 de *Lecture Notes in Computer Science*, pages 121–136. Springer-Verlag. Cité 1 fois, p. 38.

Katz, R. H. (1990). Toward a unified framework for version modeling in engineering databases. *ACM Comput. Surv.*, 22(4):375–409. Cité 1 fois, p. 76.

Kazai, G. et Lalmas, M. (2006). INEX 2005 evaluation measures. *In* Fuhr, N., Lalmas, M., Malik, S. et Kazai, G., éditeurs : *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005, Revised Selected Papers*, volume 3977 de *Lecture Notes in Computer Science*, pages 16–29. Springer. Cité 1 fois, p. 53.

Kekäläinen, J., Junkkari, M., Arvola, P. et Aalto, T. (2005). TRIX 2004 - struggling with the overlap. *In* Fuhr, N., Lalmas, M., Malik, S. et Szlávik, Z., éditeurs : *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 127–139. Springer. Cité 1 fois, p. 37.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1—2):1–224. Cité 1 fois, p. 1.

Kelly, D. et Belkin, N. J. (2001). Reading time, scrolling and interaction : exploring implicit sources of user preferences for relevance feedback. *In SIGIR '01 : Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 408–409, New York, NY, USA. ACM. Cité 1 fois, p. 64.

Kemp, C. et Ramamohanarao, K. (2002). Long-term learning for Web search engines. *In PKDD '02 : Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 263–274, London, UK. Springer-Verlag. Cité 1 fois, p. 66.

Khrouf, K. et Soulé-Dupuy, C. (2005). DocWare : Vers l'entreposage et l'analyse multidimensionnelle de documents. *In CORIA'05 : COnférence en Recherche d'Infomations et Applications*, pages 405–420. Cité 1 fois, p. 62.

Kiryakov, A., Popov, B., Terziev, I., Manov, D. et Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics : Science, Services and Agents on the World Wide Web*, 2(1):49 – 79. Cité 2 fois, p. 11 et 29.

Klink, S. (2004). Improving document transformation techniques with collaborative learned term-based concepts. *In Reading and Learning, Adaptive Content Recognition*, volume 2956 de *Lecture Notes in Computer Science*, pages 281–305. Springer. Cité 2 fois, p. 65 et 66.

Koller, D. et Sahami, M. (1997). Hierarchically classifying documents using very few words. *In ICML '97 : Proceedings of the Fourteenth International Conference on Machine Learning*, pages 170–178, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Cité 1 fois, p. 11.

Korfhage, R. R. et Chavarria Garza, H. (1982). Retrieval improvement by interaction of queries and user profiles. *In COMPSAC'82 : Sixth International Conference on Computer Software & Applications*, pages 470–475, Washington, DC, USA. IEEE. Cité 2 fois, p. 2 et 4.

Korfhage, R. R., Lin, X. et Dubin, D. S. (1995). VIRI : Visual information retrieval interfaces. *In Post-Conference Research Workshops - The 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 170–178, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Cité 1 fois, p. 90.

Kostadinov, D. (2007). *Personnalisation de l'information : une approche de gestion de profils et de reformulation de requêtes*. Thèse de doctorat, Université de Versailles Saint-Quentin en Yvelines, Versailles Saint-Quentin en Yvelines, France. Cité 2 fois, p. 65 et 66.

Kuhn, K. (2004). Problems and benefits of requirements gathering with focus groups : A case study. *International Journal of Human-Computer Interaction*, 12(3):309–325. Cité 1 fois, p. 92.

Kwok, K. L. (2005). An attempt to identify weakest and strongest queries. *In Workshop on Predicting Query Difficulty, 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Cité 1 fois, p. 88.

Kwok, K. L., Grunfeld, L. et Xu, J. H. (1997). TREC-6 English and Chinese retrieval experiments using PIRCS. *In TREC*, pages 207–214. Cité 1 fois, p. 69.

Kwon, O.-W., Kim, M.-C. et Choi, K.-S. (1994). Query expansion using domain-adapted, weighted thesaurus in an extended boolean model. *In CIKM '94 : Proceedings of the third international conference on Information and knowledge management*, pages 140–146, New York, NY, USA. ACM. Cité 2 fois, p. 8 et 11.

L

Lagus, K., Kaski, S., Honkela, T. et Kohonen, T. (1996). Browsing digital libraries with the aid of self-organizing maps. *In WWW5 : The Fifth International World Wide Web Conference*, volume Poster Proceedings, pages 71–79. Cité 1 fois, p. 90.

Lainé-Cruzel, S. (1999). ProfilDoc : Filtrer une information exploitable. *Bulletin des bibliothèques de France*, 44(5):60–64. Cité 2 fois, p. 65 et 114.

Lalmas, M. (2009). XML retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–111. Cité 1 fois, p. 36.

Lalmas, M. et Roelleke, T. (2004). Modelling vague content and structure querying in XML retrieval with a probabilistic object-relational framework. *In Proceedings of the 6th International Conference on Flexible Query Answering Systems (FQAS)*, volume 3055 de *LNCS*, pages 432–445. Springer, Berlin et al. Cité 1 fois, p. 37.

Larson, R. R. (2006). Probabilistic retrieval, component fusion and blind feedback for XML retrieval. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 225–239, New York, NY, USA. Springer-Verlag. Cité 1 fois, p. 36.

Le, T. H. D., Chevallet, J.-P. et Dong, T. B. T. (2007). Thesaurus-based query and document expansion in conceptual indexing with UMLS : Application in medical information retrieval. *In Research, Innovation and Vision for the Future, 2007 IEEE International Conference on*, pages 242 –246. Cité 2 fois, p. 8 et 11.

Lee, J. H. (1997). Analyses of multiple evidence combination. *In SIGIR '97 : Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA. ACM. Cité 4 fois, p. 22, 26, 72 et 90.

Lewis, D. D. et Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. *In SDAIR-94 : 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93. Cité 1 fois, p. 11.

Lewis, D. D., Schapire, R. E., Callan, J. P. et Papka, R. (1996). Training algorithms for linear text classifiers. *In SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306, New York, NY, USA. ACM. Cité 2 fois, p. 10 et 21.

Li, S., Xia, R., Zong, C. et Huang, C.-R. (2009). A framework of feature selection methods for text categorization. *In ACL-IJCNLP '09 : Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2*, pages 692–700, Morristown, NJ, USA. Association for Computational Linguistics. Cité 1 fois, p. 89.

Liu, H., Li, J. et Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13:51–60. Cité 1 fois, p. 89.

Liu, Q., Jiang, M. et Chen, Z. (2008). Query recommendation with tf-iqf model and popularity factor. *In FSKD '08 : Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, pages 203–207, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 66.

Liu, S., Zou, Q. et Chu, W. W. (2004). Configurable indexing and ranking for XML information retrieval. *In SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 88–95, New York, NY, USA. ACM. Cité 1 fois, p. 89.

Lu, W., Robertson, S. et MacFarlane, A. (2006). Field-weighted XML retrieval based on BM25. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 161–171, New York, NY, USA. Springer-Verlag. Cité 2 fois, p. 36 et 37.

M

Mandl, T. et Womser-Hacker, C. (2003). Linguistic and statistical analysis of the CLEF topics. *In* Peters, C., Braschler, M., Gonzalo, J. et Kluck, M., éditeurs : *CLEF*, volume 2785 de *Lecture Notes in Computer Science*, pages 505–511. Springer. Cité 1 fois, p. 88.

Mandl, T. et Womser-Hacker, C. (2005). A content independent model for context adaptation and individualization in information retrieval. *In Proceedings of the International Workshop on Context-Based Information Retrieval (CIR-05) jointly with the 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, volume 151 de *CEUR Workshop Proceedings*. CEUR-WS.org. Cité 1 fois, p. 89.

Manning, C. D., Raghavan, P. et Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. Cité 1 fois, p. 23.

Martín-Bautista, M. J., Kraft, D. H., Miranda, M. A. V., Chen, J. et Cruz, J. (2002). User profiles and fuzzy logic for web retrieval issues. *Soft Computing*, 6(5):365–372. Cité 2 fois, p. 64 et 65.

Mass, Y. et Mandelbrod, M. (2005). Component ranking and automatic query refinement for XML retrieval. *In Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 73–84. Springer. Cité 4 fois, p. 36, 37, 41 et 110.

Mass, Y. et Mandelbrod, M. (2006). Using the INEX environment as a test bed for various user models for XML retrieval. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 187–195, New York, NY, USA. Springer-Verlag. Cité 2 fois, p. 37 et 38.

McCracken, D. D., Spool, J. M. et Wolfe, R. J. (2003). *User-Centered Web Site Development : A Human-Computer Interaction Approach*. Pearson Education. Cité 1 fois, p. 116.

McGuinness, D. et van Harmelen, F. (2004). OWL Web Ontology Language, Overview, WWW Consortium, Recommendation REC-owl-features-20040210. Cité 1 fois, p. 29.

Mihajlovic, V., Ramírez, G., Westerveld, T., Hiemstra, D., Blok, H. E. et de Vries, A. (2006). TIJAH scratches INEX 2005 : Vague element selection, image search, overlap, and relevance feedback. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 72–87, New York, NY, USA. Springer-Verlag. Cité 1 fois, p. 38.

Mihalcea, R. (2002). Classifier stacking and voting for text filtering. *In TREC 2002 : The Eleventh Text REtrieval Conference*. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 89.

Mizzaro, S. et Robertson, S. (2007). Hits hits TREC : exploring IR evaluation results with network analysis. *In SIGIR '07 : Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 479–486, New York, NY, USA. ACM. Cité 1 fois, p. 88.

Montague, M. et Aslam, J. A. (2002). Condorcet fusion for improved retrieval. *In CIKM '02 : Proceedings of the eleventh international conference on Information and knowledge management*, pages 538–548, New York, NY, USA. ACM. Cité 3 fois, p. 22, 26 et 90.

Moreau, F. et Sébillot, P. (2005). Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport technique Publication interne n°1690, IRISA. Cité 1 fois, p. 107.

Mothe, J., Augé, J., Benammar, A. et Hubert, G. (2002a). Deliverable : D.3.1.8. (Pre-prototyping), IRAIA IST-1999-10602. Rapport technique Rapport de contrat, D.3.1.8. IRAIA, EC. Cité 4 fois, p. 12, 33, 160 et 162.

Mothe, J., Augé, J. et Hubert, G. (2002b). Deliverable : D.3.1.7 System Specification, IRAIA IST-1999-10602. Rapport technique Rapport de contrat, D.3.1.7 IRAIA, EC. Cité 5 fois, p. 12, 33, 160, 162 et 163.

Mothe, J., Chrisment, C., Dousset, B. et Alaux, J. (2003). DocCube : multi-dimensional visualisation and exploration of large document sets. *J. Am. Soc. Inf. Sci. Technol.*, 54(7):650–659. Cité 1 fois, p. 62.

Mothe, J. et Tanguy, L. (2005). Linguistic features to predict query difficulty - a case study on previous TREC campaigns. *In Workshop on Predicting Query Difficulty, 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Cité 1 fois, p. 88.

Mylonas, P., Vallet, D., Castells, P., Fernández, M. et Avrithis, Y. (2008). Personalized information retrieval based on context and ontological knowledge. *Knowl. Eng. Rev.*, 23(1):73–100. Cité 1 fois, p. 66.

N

Naderi, H., Rumpler, B. et Pinon, J.-M. (2007). An efficient collaborative information retrieval system by incorporating the user profile. *In AMR'06 : Proceedings of the 4th international conference on Adaptive multimedia retrieval*, volume 4398 de *Lecture Notes In Computer Science*, pages 247–257, Berlin, Heidelberg. Springer-Verlag. Cité 1 fois, p. 66.

Naderi, H., Rumpler, B. et Pinon, J.-m. (2008). A graph-based profile similarity calculation method for collaborative information retrieval. *In SAC '08 : Proceedings of the 2008 ACM symposium on Applied computing*, pages 1127–1131, New York, NY, USA. ACM. Cité 1 fois, p. 66.

Nagypál, G. (2005). Improving information retrieval effectiveness by using domain knowledge stored in ontologies. *In On the Move to Meaningful Internet Systems 2005 : OTM Workshops, Ph.D. Student Symposium*, volume 3762 de *Lecture Notes In Computer Science*, pages 780–789. Springer-Verlag. Cité 1 fois, p. 8.

Nanas, N., Uren, V. et De Roeck, A. (2003). Building and applying a concept hierarchy representation of a user profile. *In SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 198–204, New York, NY, USA. ACM. Cité 1 fois, p. 65.

Nielsen, J. et Mack, R. L., éditeurs (1994). *Usability inspection methods*. John Wiley & Sons, Inc., New York, NY, USA. Cité 1 fois, p. 92.

Nielsen, J. et Molich, R. (1990). Heuristic evaluation of user interfaces. *In CHI '90 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256, New York, NY, USA. ACM. Cité 1 fois, p. 91.

O

Ogilvie, P. et Callan, J. (2003). Combining document representations for known-item search. *In SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 143–150, New York, NY, USA. ACM. Cité 2 fois, p. 37 et 115.

Ogilvie, P. et Callan, J. (2006). Parameter estimation for a simple hierarchical generative model for XML retrieval. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 211–224, New York, NY, USA. Springer-Verlag. Cité 2 fois, p. 36 et 37.

P

Palacio, D., Cabanac, G., Sallaberry, C. et Hubert, G. (2010a). Cadre d'évaluation de systèmes de recherche d'information géographique : apport de la combinaison des dimensions spatiale, temporelle et thématique (regular paper). *In INFORSID 2010 : Congrès Informatique des Organisations et Systèmes d'Information et de Décision*, pages 245–260, http ://inforsid.irit.fr. INFORSID. Cité 6 fois, p. 33, 34, 61, 131, 159 et 161.

Palacio, D., Cabanac, G., Sallaberry, C. et Hubert, G. (2010b). Measuring Effectiveness of Geographic IR Systems in Digital Libraries : Evaluation Framework and Case Study. *In* Lalmas, M., Jose, J., Rauber, A., Sebastiani, F. et Frommholz, I., éditeurs : *ECDL'10 : Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, volume 6273 de *Lecture Notes In Computer Science*, pages 340–351. Springer. Cité 6 fois, p. 33, 34, 61, 131, 159 et 161.

Pass, G., Chowdhury, A. et Torgeson, C. (2006). A picture of search. *In InfoScale '06 : Proceedings of the 1st international conference on Scalable information systems*, page 1, New York, NY, USA. ACM. Cité 2 fois, p. 78 et 80.

Pehcevski, J., Thom, J. A. et Tahaghoghi, S. M. M. (2005). Hybrid XML retrieval revisited. *In Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 153–167. Springer. Cité 1 fois, p. 89.

Pinel-Sauvagnat, K., Hubert, G., Boughanem, M. et Mothe, J. (2003). IRIT at INEX 2003. *In Inex 2003 : International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 142–148. Cité 4 fois, p. 38, 43, 61 et 161.

Piwowarski, B. et Gallinari, P. (2005). A bayesian framework for XML information retrieval : Searching and learning with the INEX collection. *Inf. Retr.*, 8(4):655–681. Cité 1 fois, p. 36.

Plaisant, C., Fekete, J.-D. et Grinstein, G. (2008). Promoting insight-based evaluation of visualizations : From contest to benchmark repository. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):120–134. Cité 2 fois, p. 92 et 93.

Plovnick, R. M. et Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology : a pilot study. *Journal of medical Internet research*, 6(3). Cité 1 fois, p. 8.

Popovici, E., Ménier, G. et Marteau, P.-F. (2007). Interprétation vague des contraintes structurelles pour la ri dans des corpus de documents XML : Évaluation d'une méthode approchée de ri structurée. *Document numérique*, 10(1). Cité 2 fois, p. 37 et 38.

Pouliquen, B., Delamarre, D. et Le Beux, P. (2002). Indexation de textes médicaux par extraction de concepts, et ses utilisations. *In JADT'2002 : 6th International Conference on the Statistical Analysis of Textual Data*, volume 2, pages 617–628. Cité 2 fois, p. 11 et 28.

Pretschner, A. et Gauch, S. (1999). Ontology based personalized search. *In ICTAI '99 : Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, page 391, Washington, DC, USA. IEEE Computer Society. Cité 2 fois, p. 65 et 66.

Psarras, I. et Jose, J. M. (2006). A system for adaptive information retrieval. *In* Wade, V. P., Ashman, H. et Smyth, B., éditeurs : *AH 2006 : 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 4018 de *Lecture Notes in Computer Science*, pages 313–317. Springer. Cité 1 fois, p. 64.

Q

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. Cité 1 fois, p. 118.

R

Raggett, D., Le Hors, A. et Jacobs, I. (1998). HTML 4.0 Specification, W3C Recommendation. Cité 1 fois, p. 35.

Raghavan, V., Bollmann, P. et Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229. Cité 1 fois, p. 52.

Raghavan, V. V. et Sever, H. (1995). On the reuse of past optimal queries. *In SIGIR '95 : Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 344–350, New York, NY, USA. ACM. Cité 2 fois, p. 65 et 66.

Raja, F., Keikha, M., Rahgozar, M. et Oroumchian, F. (2007). Effectiveness of rich document representation in XML retrieval. *In* Evans, D., Furui, S. et Soulé-Dupuy, C., éditeurs : *RIAO 2007 : 8th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications)*. CID. Cité 1 fois, p. 36.

Ralalason, B. J. V. (2010). *Représentation multi-facettes des documents pour leur accès sémantique*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France. Cité 4 fois, p. 30, 34, 156 et 163.

Ravat, F., Teste, O., Tournier, R. et Zurfluh, G. (2010). Finding an application-appropriate model for XML data warehouses. *Information Systems*, 35(6):662–687. Cité 1 fois, p. 62.

Revuri, S., Upadhyaya, S. R. et Kumar, P. S. (2007). Using domain ontologies for efficient information retrieval. *In COMAD 2006 : 13th International Conference on Management of Data*. Cité 1 fois, p. 8.

Robertson, G. G., Mackinlay, J. D. et Card, S. K. (1991). Cone trees : animated 3D visualizations of hierarchical information. *In CHI '91 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 189–194, New York, NY, USA. ACM. Cité 1 fois, p. 91.

Robertson, S. E. (1977). The probabilistic character of relevance. *Inf. Process. Manage.*, 13(4):247–251. Cité 1 fois, p. 1.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gatford, M. et Payne, A. (1995). Okapi at TREC-4. *In TREC-4 : The 4th Text REtrieval Conference*, pages 73–97. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 89.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M. et Willet, P. (1998). Okapi at TREC-7 : Automatic ad hoc, filtering, VLC and interactive track. *In TREC-7 : The 7th Text REtrieval Conference*, pages 253–264. National Institute of Standards and Technology (NIST). Cité 1 fois, p. 110.

Rocchio, J. (1971). *Relevance Feedback in Information Retrieval*, pages 313–323. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. Cité 4 fois, p. 2, 10, 21 et 69.

Rohrer, R. M., Sibert, J. L. et Ebert, D. S. (1999). A shape-based visual interface for text retrieval. *IEEE Comput. Graph. Appl.*, 19(5):40–46. Cité 1 fois, p. 91.

Rooney, N., Patterson, D., Galushka, M. et Dobrynin, V. (2006). A relevance feedback mechanism for cluster-based retrieval. *Inf. Process. Manage.*, 42(5):1176–1184. Cité 1 fois, p. 64.

Rosenfeld, L. et Morville, P. (1998). *Information architecture for the World Wide Web*. O'Reilly & Associates, Inc., Sebastopol, CA, USA. Cité 1 fois, p. 115.

S

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. Cité 2 fois, p. 1 et 17.

Salton, G., Allan, J. et Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *In SIGIR '93 : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58, New York, NY, USA. ACM. Cité 1 fois, p. 3.

Salton, G. et Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523. Cité 1 fois, p. 24.

Salton, G. et Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297. Cité 1 fois, p. 64.

Salton, G., Wong, A. et Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620. Cité 1 fois, p. 14.

Sandu Popa, I., Zeitouni, K., Gardarin, G., Nakache, D. et Metais, E. (2007). Text categorization for multi-label documents and many categories. *In CBMS '07 : Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems*, pages 421–426, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 11.

Saracevic, T. (1997). The stratified model of information retrieval interaction : Extension and applications. *In Proceedings of the American Society for Information Science*, volume 34, pages 313–327. Cité 1 fois, p. 2.

Sauvagnat, K., Boughanem, M. et Chrisment, C. (2006). Answering content and structure-based queries on XML documents using relevance propagation. *Inf. Syst.*, 31(7):621–635. Cité 1 fois, p. 37.

Schapire, R. E., Singer, Y. et Singhal, A. (1998). Boosting and Rocchio applied to text filtering. *In SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 215–223, New York, NY, USA. ACM. Cité 1 fois, p. 11.

Schütze, H., Hull, D. A. et Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. *In SIGIR '95 : Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 229–237, New York, NY, USA. ACM. Cité 1 fois, p. 11.

Schweighofer, E. et Geist, A. (2007). Legal query expansion using ontologies and relevance feedback. *In* Casanovas, P., Biasiotti, M. A., Francesconi, E. et Sagri, M.-T., éditeurs : *LOAIT '07 : Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques*, volume 321 de *CEUR Workshop Proceedings*, pages 149–160. CEUR-WS.org. Cité 1 fois, p. 8.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47. Cité 1 fois, p. 11.

Selberg, E. et Etzioni, O. (1998). Experiments with Collaborative Index Enhancement. Rapport technique UW-CSE-98-06-01, Department of Computer Science and Engineering, University of Washington, Seattle, WA. Cité 2 fois, p. 65 et 66.

Shi, X. et Yang, C. C. (2007). Mining related queries from web search engine query logs using an improved association rule mining model. *J. Am. Soc. Inf. Sci. Technol.*, 58(12):1871–1883. Cité 1 fois, p. 66.

Shiri, A., Ruecker, S., Rossello, X., Bouchard, M. et Mehta, P. (2007). Development of a thesaurus-enhanced visual interface for multilingual digital libraries. *In 35th Annual Conference of the Canadian Association for Information Science, Information Sharing in a Fragmented World : Crossing Boundaries*. CAIS. Cité 2 fois, p. 10 et 11.

Shneiderman, B. (1997). *Designing the User Interface : Strategies for Effective Human-Computer Interaction*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. Cité 1 fois, p. 92.

Shneiderman, B. et Plaisant, C. (2004). *Designing the User Interface : Strategies for Effective Human-Computer Interaction (4th Edition)*. Pearson Addison Wesley. Cité 1 fois, p. 92.

Sieg, A., Mobasher, B. et Burke, R. (2007). Web search personalization with ontological user profiles. *In CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525–534, New York, NY, USA. ACM. Cité 1 fois, p. 65.

Sigurbjörnsson, B. et Kamps, J. (2006). The effect of structured queries and selective indexing on XML retrieval. *In Proceedings of the Advances in XML Information Retrieval and Evaluation : 4th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 de *Lecture Notes in Computer Science*, pages 104–118, New York, NY, USA. Springer-Verlag. Cité 1 fois, p. 37.

Sigurbjörnsson, B., Kamps, J. et de Rijke, M. (2005). Mixture models, overlap, and structural hints in XML element retrieval. *In Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 196–210. Springer. Cité 3 fois, p. 37, 41 et 89.

Singhal, A. (2001). Modern information retrieval : A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42. Cité 1 fois, p. 14.

Singhal, A., Buckley, C. et Mitra, M. (1996). Pivoted document length normalization. *In SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, New York, NY, USA. ACM. Cité 1 fois, p. 89.

Smith, S. L. et Mosier, J. N. (1986). Design guidelines for designing user interface software. Rapport technique ESD-TR-86-278, The MITRE Corporation, Bedford, MA. Cité 1 fois, p. 92.

Smucker, M. D., Allan, J. et Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *In CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632, New York, NY, USA. ACM. Cité 1 fois, p. 27.

Speretta, M. et Gauch, S. (2005). Personalized search based on user search histories. *In WI '05 : Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 622–628, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 66.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21. Cité 1 fois, p. 17.

Stefani, A. et Strappavara, C. (1998). Personalizing access to Web sites : The SiteIF project. *In HYPERTEXT '98 : 2nd Workshop on Adaptive Hypertext and Hypermedia*. Cité 1 fois, p. 65.

Stuckenschmidt, H., van Harmelen, F., de Waard, A., Scerri, T., Bhogal, R., van Buel, J., Crowlesmith, I., Fluit, C., Kampman, A., Broekstra, J. et van Mulligen, E. (2004). Exploring large document repositories with RDF technology : The DOPE project. *IEEE Intelligent Systems*, 19:34–40. Cité 1 fois, p. 10.

T

Tamine, L., Boughanem, M. et Zemirli, W. N. (2007). Exploiting multi-evidence from multiple users interests to personalizing information retrieval. *In* Badr, Y., Chbeir, R. et Pichappan, P., éditeurs : *ICDIM '07 : IEEE International Conference on Digital Information Management*, pages 7–12, http ://computer.org/cspress. IEEE Engineering Management Society. Cité 1 fois, p. 66.

Tannier, X. (2006). Recherche d'information dans les documents XML. Rapport technique 2006-400-007, Ecole Nationale Supérieure des Mines de Saint-Etienne. Cité 1 fois, p. 36.

Taylor, M., Guiver, J., Robertson, S. et Minka, T. (2008). SoftRank : optimizing non-smooth rank metrics. *In WSDM '08 : Proceedings of the international conference on Web search and web data mining*, pages 77–86, New York, NY, USA. ACM. Cité 2 fois, p. 89 et 93.

Theobald, M., Broschart, A., Schenkel, R., Solomon, S. et Weikum, G. (2007). TopX - AdHoc Track and Feedback Task. *In* Fuhr, N., Lalmas, M. et Trotman, A., éditeurs : *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Revised and Selected Papers*, volume 4518 de *Lecture Notes in Computer Science*, pages 233–242. Springer. Cité 1 fois, p. 38.

Theobald, M., Schenkel, R. et Weikum, G. (2006). TopX and XXL at INEX 2005. *In* Fuhr, N., Lalmas, M., Malik, S. et Kazai, G., éditeurs : *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Revised Selected Papers*, volume 3977 de *Lecture Notes in Computer Science*, pages 282–295. Springer. Cité 1 fois, p. 37.

Thomas, J. et Cook, K., éditeurs (2005). *Illuminating the Path : The Research and Development Agenda for Visual Analytics*. IEEE CS Press. Cité 2 fois, p. 92 et 113.

Tomassen, S. L., Gulla, J. A. et Strasunskas, D. (2006). Document space adapted ontology : Application in query enrichment. *In* Kop, C., Fliedl, G., Mayr, H. C. et Métais, E., éditeurs : *Natural Language Processing and Information Systems, 11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006*, volume 3999 de *Lecture Notes in Computer Science*, pages 46–57. Springer. Cité 1 fois, p. 8.

Trotman, A. et Sigurbjörnsson, B. (2005). Narrowed Extended XPath I (NEXI). *In* Fuhr, N., Lalmas, M., Malik, S. et Szlávik, Z., éditeurs : *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Revised Selected Papers*, volume 3493 de *Lecture Notes in Computer Science*, pages 16–40. Springer. Cité 3 fois, p. 36, 39 et 52.

Trousse, B., Jaczynski, M. et Kanawati, R. (1999). Using user behaviour similarity for recommendation computation : the Broadway approach. *In Proceedings of the HCI International '99 (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction*, volume 2, pages 85–89, Hillsdale, NJ, USA. L. Erlbaum Associates Inc. Cité 1 fois, p. 65.

Tuominen, J., Kauppinen, T., Viljanen, K. et Hyvönen, E. (2009). Ontology-based query expansion widget for information retrieval. *In Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009)*. Cité 1 fois, p. 8.

U

Ukkonen, E. (1985). Algorithms for approximate string matching. *Inf. Control*, 64(1-3):100–118. Cité 1 fois, p. 31.

V

Vallet, D., Fernández, M. et Castells, P. (2005). An ontology-based information retrieval model. *In The Semantic Web : Research and Applications*, volume 3532 de *Lecture Notes in Computer Science*, pages 455–470. Cité 1 fois, p. 10.

van Heijst, G., Schreiber, A. T. et Wielinga, B. J. (1997). Using explicit ontologies in KBS development. *Int. J. Hum.-Comput. Stud.*, 46(2-3):183–292. Cité 1 fois, p. 29.

van Rijsbergen, C. J. (1975). *Information retrieval*. Butterworths, London ; Boston :. Cité 1 fois, p. 17.

van Zwol, R. (2006). B^3-SDR and effective use of structural hints. *In* Fuhr, N., Lalmas, M., Malik, S. et Kazai, G., éditeurs : *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Revised Selected Papers*, volume 3977 de *Lecture Notes in Computer Science*, pages 146–160. Springer. Cité 1 fois, p. 37.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA. Cité 1 fois, p. 10.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A. et Ciravegna, F. (2002). MnM : Ontology driven semi-automatic and automatic support for semantic markup. *In EKAW '02 : Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web*, volume 2473 de *Lecture Notes in Computer Science*, pages 379–391, London, UK. Springer-Verlag. Cité 1 fois, p. 11.

Vinot, R. (2004). *Classification automatique de textes dans des catégories non thématiques*. Thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris, France. Cité 1 fois, p. 10.

Vinot, R. et Yvon, F. (2003). Improving Rocchio with weakly supervised clustering. *In* Lavrac, N., Gamberger, D., Todorovski, L. et Blockeel, H., éditeurs : *Machine Learning : ECML 2003, 14th European Conference on Machine Learning*, volume 2837 de *Lecture Notes in Computer Science*, pages 456–467. Springer. Cité 1 fois, p. 10.

Vu, H.-T. et Gallinari, P. (2008). Analyse de la robustesse des algorithmes de méta-recherche discriminante. *In CORIA*, pages 87–102. Cité 1 fois, p. 90.

W

Wharton, C., Rieman, J., Lewis, C. et Polson, P. (1994). The cognitive walkthrough method : a practitioner's guide. pages 105–140. Cité 1 fois, p. 91.

White, R., Ruthven, I. et Jose, J. M. (2002). The use of implicit evidence for relevance feedback in Web retrieval. *In Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, volume 2291 de *Lecture Notes in Computer Science*, pages 93–109, London, UK. Springer-Verlag. Cité 1 fois, p. 64.

Wollersheim, D. et Rahayu, J. W. (2005). Ontology based query expansion framework for use in medical information systems. *International Journal of Web Information Systems*, 1(2):101–115. Cité 1 fois, p. 8.

Wu, S. (2009). Applying statistical principles to data fusion in information retrieval. *Expert Syst. Appl.*, 36(2):2997–3006. Cité 1 fois, p. 90.

Wu, S. et McClean, S. (2006). Improving high accuracy retrieval by eliminating the uneven correlation effect in data fusion. *J. Am. Soc. Inf. Sci. Technol.*, 57(14):1962–1973. Cité 1 fois, p. 90.

X

Xia, F., Liu, T.-Y., Wang, J., Zhang, W. et Li, H. (2008). Listwise approach to learning to rank : theory and algorithm. *In ICML '08 : Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, New York, NY, USA. ACM. Cité 1 fois, p. 89.

Xu, J. et Croft, W. B. (1996). Query expansion using local and global document analysis. *In SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA. ACM. Cité 1 fois, p. 69.

Xu, J. et Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112. Cité 2 fois, p. 69 et 70.

Y

Yang, Y. (1994). Expert network : effective and efficient learning from human decisions in text categorization and retrieval. *In SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22, New York, NY, USA. Springer-Verlag New York, Inc. Cité 1 fois, p. 11.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1(1-2):69–90. Cité 1 fois, p. 20.

Yom-Tov, E., Fine, S., Carmel, D., Darlow, A. et Amitay, E. (2004). Improving document retrieval according to prediction of query difficulty. *In TREC 2004 : Text REtrieval Conference*. National Institute of Standards and Technology. Cité 1 fois, p. 88.

Z

Zamir, O. et Etzioni, O. (1999). Grouper : a dynamic clustering interface to Web search results. *Computer Networks*, 31(11-16):1361–1374. Cité 1 fois, p. 90.

Zargayouna, H. et Salotti, S. (2004). Mesure de similarité sémantique pour l'indexation de documents semi-structurés. *In IC'2004 : 15e journées francophones d'Ingénierie des connaissances*. Cité 1 fois, p. 11.

Zayani, C. A., Péninou, A., Canut, M.-F. et Sèdes, F. (2009). An adaptation approach : Query enrichment by user profile. 4879:351–361. Cité 1 fois, p. 66.

Zhang, X. et Heng, X. (2008). A retrieval model based on bayesian network for XML documents. *In FSKD '08 : Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 02, pages 91–95, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 36.

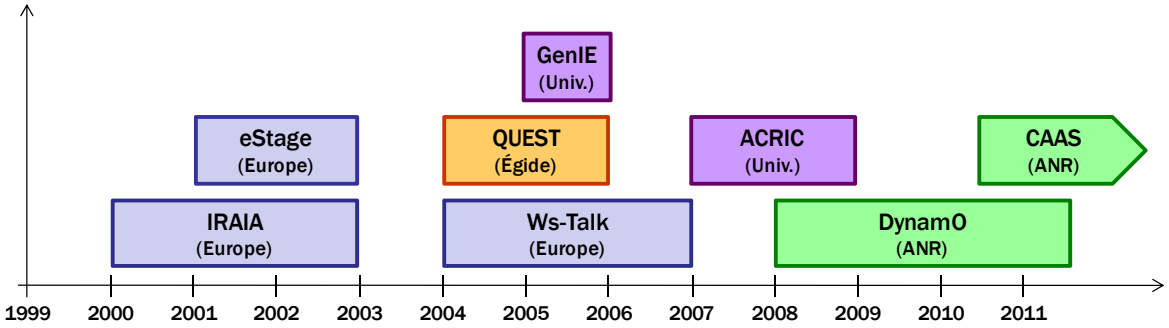
Zhu, Y. et Gruenwald, L. (2005). Query expansion using web access log files. *In* Andersen, K. V., Debenham, J. K. et Wagner, R., éditeurs : *DEXA*, volume 3588 de *Lecture Notes in Computer Science*, pages 686–695. Springer. Cité 2 fois, p. 64 et 66.

Zhuang, Z. et Cucerzan, S. (2006). Re-ranking search results using query logs. *In CIKM '06 : Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 860–861, New York, NY, USA. ACM. Cité 1 fois, p. 66.

Annexe : Encadrement et diffusion scientifique

1 Projets

Cette section synthétise les projets qui concernent mes activités de recherche.



2010 – 2012 : CAAS (Contextual Analysis and Adaptive Search)

Cadre : PROJET ANR

Financement : 440000 euros

Implication personnelle prévue : *21 h/m*

Partenaires :

- LIA, Université d'Avignon et des Pays de Vaucluse
- CLLE-ERSS, Université Toulouse II
- IRIT/SIG, Université Paul Sabatier

Objectif : L'objectif du projet CAAS est de proposer des solutions pour l'identification de contexte par la caractérisation des requêtes, des documents et des attentes des utilisateurs. Il s'agit également de caractériser les SRI afin de trouver le SRI adapté à un contexte donné.

Contribution : Ce projet est en phase de démarrage, son montage m'a permis de renforcer mes compétences en matière de constitution d'une réponse à un appel d'offre de projet ANR. J'apporte à ce projet mes compétences scientifiques en matière de d'identification de

contexte de RI et d'adaptation de système à un contexte présentées dans le chapitre 4 de ce mémoire. Ce projet supportera la poursuite de ces travaux menés précédemment lors du projet ACRIC. Je vais plus particulièrement prendre en charge les propositions, l'organisation et le suivi des lots concernant l'analyse de la variété des systèmes.

2008 – 2011 : DynamO (Dynamic Ontology for information retrieval)

Cadre : PROJET ANR

Financement : 1,2 millions d'euros

Implication personnelle : *14 h/m*

Partenaires :

- ACTIA
- ARTAL Technologies
- Préhistoire et Technologies, Université Paris X
- LaLIC, Université Paris IV
- IRIT, Université Paul Sabatier

Objectif : L'objectif principal du projet DynamO est de concevoir une approche méthodologique et un ensemble d'outils logiciels qui prennent ainsi en compte la construction et la maintenance de ressources ontologiques à partir de documents et l'utilisation de ces ressources pour une indexation sémantique facilitant la recherche d'informations.

Contribution : La participation de notre groupe s'inscrit dans plusieurs lots et plus particulièrement celui dédié à la réalisation des spécifications des outils d'annotation dynamique et de recherche dans les textes. Notre groupe est également impliqué dans le développement logiciel des solutions relatives à la recherche d'information sémantique. Dans le cadre de ce projet, je suis co-responsable du lot dédié aux spécifications des outils. À ce titre, je gère l'avancement du lot sur les aspects scientifiques et administratifs comme la rédaction des documents de spécification (Hernandez *et al.*, 2009; Dudognon *et al.*, 2010a) ainsi que les comptes-rendus d'avancement de projet lors des réunions plénières. De plus, j'ai assuré l'encadrement des activités de recherche liées à ce projet d'un étudiant en thèse (Ralalason, 2010) ainsi que le suivi des activités de développement de deux agents contractuels et d'un étudiant en stage de Master Professionnel. Les travaux réalisés dans le cadre de ce projet, qui relatifs au chapitre 1 de ce mémoire, ont donné lieu à des publications dans deux conférences nationales (Hubert *et al.*, 2009a; Dudognon *et al.*, 2010b). J'ai apporté à ce projet mes compétences scientifiques en recherche d'information sémantique que j'ai complétées au contact des partenaires du projet dans le but de constituer de nouvelles propositions de projets pour supporter la poursuite de nos travaux en recherche d'information sémantique. J'ai complété mes connaissances en en matière de réponse à un appel à projet, plus précisément dans le cadre ANR, que j'ai pu mettre à profit dans la rédaction du projet accepté CAAS en phase de démarrage. J'ai pu également renforcer mes compétences en matière de gestion de l'avancement d'un projet notamment vis-à-vis de l'ANR et de gestion du financement des travaux de recherche que je pourrai également réinvestir dans le projet CAAS. Dans l'environnement de ce projet, j'ai pu appréhender les conventions CIFRE par la constitution d'une proposition de thèse de ce type avec l'un des partenaires industriels du projet.

2007 – 2008 : ACRIC (Analyse Canonique et Recherche d'Information Contextuelle)

Cadre : PROJET DU PROGRAMME PLURI-FORMATIONS DE RECHERCHE EN MATHÉMATIQUES ET INFORMATIQUE DE TOULOUSE

Financement : 6000 euros

Partenaires :

- Institut de Mathématiques de Toulouse (IMT), équipe Statistiques et Probabilités (LSP)
- Institut de Recherche en informatique de Toulouse (IRIT), équipe SIG-EVI

Objectif : Le projet ACRIC s'est intéressé tout d'abord à la mise en relation des critères linguistiques et des performances des systèmes au moyen de techniques statistiques telles que l'analyse canonique. Il s'est intéressé à l'étude plus approfondie de la notion de contexte en RI.

Contribution : Dans le cadre de ce projet, j'ai proposé et supervisé les travaux autour à la sélection du système efficace répondant à des éléments contextuels donnés décrits au chapitre 4 de ce mémoire. La coordination et la gestion de la complémentarité d'équipes appartenant à différentes disciplines ont constitué une expérience qui pourra être réinvestie dans le projet CAAS. Les travaux réalisés dans le cadre de ce projet ont donné lieu à des publications dans une revue nationale (Hubert et Mothe, 2006), deux conférences internationales (Hubert et Mothe, 2007a; Hubert *et al.*, 2007c) et trois conférences nationales (Englmeier *et al.*, 2006; Hubert *et al.*, 2007a,b).

2004 – 2006 : Ws-Talk (Web services communicating in the language of their user community)

Cadre : PROJET EUROPÉEN IST 6E PROGRAMME CADRE

Financement : 1,05 millions d'euros

Implication personnelle : *8 h/m*

Partenaires :

- Universidad Diego Portales, Chili
- Universidad de Talca, Chili
- Soluciones S. A., Chili
- Royal Holloway & Bedford New College, Royaume-Uni
- Queen's PlaceTypeplaceUniversity of PlaceNameBelfast, Royaume-Uni
- LuckyEye A. S., Turquie
- Foundation for Research and Technology – placeHellas, Grèce
- Archimedia S. A., Grèce
- Akra Gmbh, Allemagne
- Université Paul Sabatier – IRIT/SIG, France

Objectif : Le but du projet Ws-Talk a été de combler le fossé qui existe entre le langage naturel quotidien utilisé pour décrire les besoins d'informations et celui utilisé pour manipuler

nombre de services Web comme par exemple le langage XML. L'objectif a été le développement de systèmes permettant de traduire une requête du client en forme compréhensible par une machine, de localiser les services Web appropriés, de les interroger et de renvoyer les résultats à l'utilisateur.

Contribution : La participation de notre groupe a concerné la spécification et le développement d'un outil de conception de vocabulaire contrôlé. En se basant sur une analyse d'un ensemble de documents XML déterminés par l'utilisateur comme étant représentatifs de son domaine professionnel, cet outil a eu pour objectif de construire un vocabulaire contrôlé pour ce domaine d'application. Ce projet s'inscrit dans mes activités de recherche relative à la RI XML détaillées au chapitre 2 de ce mémoire. Dans le cadre de ce projet, j'ai pris en charge la définition et le développement d'un module d'indexation de documents XML. J'ai également assuré le suivi des activités de développement de deux agents contractuels recrutés dans le cadre de ce projet. Ce projet a consolidé mes acquis en matière de gestion d'un projet européen. J'ai acquis de l'expérience dans la conception d'un projet en lots cohérents que j'ai pu réutiliser dans les propositions de projet suivantes. Les travaux réalisés dans le cadre de ce projet ont donné lieu à des publications dans deux workshops internationaux (Hubert, 2005, 2007) et une conférence nationale (Hubert *et al.*, 2005). De plus, j'ai défini et encadré un stage de Master Recherche en lien avec ce projet (Hamza, 2007).

2005 : GenIE (Genomic Information Extraction)

Cadre : BQR UNIVERSITÉ PAUL SABATIER

Partenaires :

- Institut de Mathématiques de Toulouse (IMT), équipe Statistiques et Probabilités (LSP)
- IRIT/SIG Université Paul Sabatier
- Cayla, Toulouse
- Libragen, Toulouse

Objectif : L'objectif du projet GenIE a été de dresser l'état de l'art des travaux existants pour le traitement d'information liée à la génomique et de dégager les axes de recherches possibles restant à explorer.

Contribution : Dans le cadre de ce projet, j'ai participé à l'étude des travaux existants et à la rédaction du rapport final. Ce projet m'a ouvert aux problématiques et travaux pour le traitement d'information dans le domaine de la génomique. Il s'inscrit dans mes activités de recherche sur la prise en compte du domaine dans la RI présentées au chapitre 1 de ce mémoire. La constitution d'un consortium pluridisciplinaire a constitué l'une des compétences que j'ai acquise dans ce projet de même que les facettes que regroupe un projet de type BQR.

2004 – 2005 : QUEST (QUEry reformulation for STructured document retrieval)

Cadre : PROJET ÉGIDE ALLIANCE – GRANDE-BRETAGNE

Partenaires :

- Queen Mary Information Retrieval (QMIR), Queen Mary, University of London
- IRIT/SIG Université Paul Sabatier

Objectif : L'objectif du projet QUEST a été de concevoir, implanter et évaluer des mécanismes pour la reformulation de requêtes pour une recherche de documents structurés efficace. Ces mécanismes combinent contenu, structure et connaissance sémantique pour aider les utilisateurs à accéder à des collections volumineuses de documents XML, langage de structuration de données adopté par le W3C.

Contribution : Ce projet s'inscrit dans les travaux de recherche en RI XML détaillés dans le chapitre 2 de ce mémoire. Dans le cadre de ce projet, j'ai ainsi proposé et développé un prototype de SRI XML en m'appuyant sur les contributions apportées dans les projets IRAIA et eStage. L'efficacité des mécanismes développés sur de grandes collections étant une orientation du projet, j'ai également pris en charge la participation à différentes éditions du programme d'évaluation INEX regroupant la définition de requêtes, la soumission de résultats (« runs »), la définition de jugements de pertinence et l'évolution du prototype de SRI XML. Les travaux réalisés dans le cadre de ce projet ont donné lieu à des publications dans deux workshops internationaux (Hubert, 2005, 2006). L'expérience acquise a été dans l'évaluation des systèmes que j'ai mise à profit dans mes travaux sur les VIRI présentés au chapitre 4 de ce mémoire et dans ma collaboration récente avec l'équipe T2I du laboratoire LIUPPA de l'université de Pau et des Pays de l'Adour (Palacio *et al.*, 2010b,a). Elle a également concerné le cadre Égide que je réinvestis pour le financement d'une thèse en co-tutelle.

2001 – 2002 : eStage (A New Stage for the Cultural Heritage in European Puppetry)

Cadre : PROJET EUROPÉEN IST-2000-28314, 5E PROGRAMME CADRE

Coût total : 197000 euros

Implication personnelle : *4 h/m*

Partenaires :

- Deutsches Institut Fuer Wirtschaftsforschubg E. V., Allemagne
- Lemonlabs GmbH, Allemagne
- Theater Waidspeicher, Allemagne
- Université Paul Sabatier, IRIT/SIG, France

Objectif : Le projet eStage s'est appuyé sur l'infrastructure développée dans le projet IRAIA pour produire un nouveau service dédié au domaine culturel européen des marionnettes.

Contribution : À l'instar du projet IRAIA, le rôle de notre groupe dans le projet eStage a concerné l'indexation de la collection. En m'appuyant sur les solutions validées dans le projet IRAIA, j'ai pris en charge la proposition et le suivi de développement d'une approche d'indexation de document hétérogènes suivant des hiérarchies de concepts. Ce projet a consolidé mes connaissances en matière de gestion d'un projet européen que j'ai pu mettre à profit dans les projets suivants notamment le projet Ws-Talk. Les travaux réalisés dans le

cadre de ce projet s'inscrivent dans la thématique du chapitre 1 de ce mémoire. Ils ont donné lieu à une publication dans un workshop international (Benammar *et al.*, 2002a).

2000 – 2002 : IRAIA Getting Orientation in Complex Information Spaces as an Emergent Behaviour of Autonomous Information Agents)

Cadre : PROJET EUROPÉEN IST-1999-1062, 5E PROGRAMME CADRE

Coût total : 1,84 millions d'euros

Implication personnelle : *8 h/m*

Partenaires :

- Reticular Systems Inc., États-Unis
- IFO Institut Fuer Wirtschaftsforschubg E. V., Allemagne
- Iteratec GmbH, Allemagne
- Queen's University of Belfast, Royaume-Uni
- Foundation for Research and Technology – Hellas, Grèce
- Université Paul Sabatier, IRIT/SIG, France

Objectif : Le projet européen IRAIA a eu comme objectif à fournir aux utilisateurs un accès aisé à des informations telles que des documents textuels et des séries temporelles dans le domaine économique.

Contribution : Le rôle de notre groupe a concerné principalement l'indexation de la collection de documents dans un environnement multilingue dans le cadre du lot « analyse automatique de texte ». J'ai plus particulièrement proposé et supervisé des solutions pour la catégorisation automatique de document présentées dans le chapitre 1 de ce mémoire. Dans ce projet, j'ai également initié les travaux sur la prise en compte de l'utilisateur dans le processus de RI décrits dans le chapitre 3 de ce mémoire. Outre la rédaction des livrables correspondant au lot «analyse automatique de texte » (Mothe *et al.*, 2002a,b), j'ai assuré l'encadrement des activités de recherche liées à ce projet d'un étudiant en thèse (Benammar, 2003) et le suivi du développement du prototype de catégorisation qu'il a réalisé. Fort d'une expérience professionnelle dans le milieu industriel dans plusieurs projets, j'ai également supervisé le développement logiciel de la solution d'indexation, réalisé par un agent contractuel recruté pour ce projet. Les travaux réalisés dans le cadre de ce projet ont donné lieu à des publications dans une revue nationale (Benammar *et al.*, 2003), 2 conférences internationales (Hubert *et al.*, 2001; Benammar *et al.*, 2002b), et trois conférences nationales (Augé *et al.*, 2001; Benammar *et al.*, 2001; Augé *et al.*, 2003). L'expérience acquise a été dans la gestion d'un projet européen notamment du point de vue de la coordination entre les partenaires de différents pays. Elle l'a été également dans la gestion du financement que j'ai pu réinvestir dans le projet Ws-Talk. De plus, j'ai poursuivi ensuite les propositions initiées dans ce projet concernant la gestion de profils d'utilisateurs notamment au travers de la proposition et l'encadrement de deux stages de Master Recherche (Al Sabbagh, 2005; El Makssoud, 2010).

2 Publications

Cette section récapitule mes publications et les encadrements de thèses et de masters recherche auxquels j'ai participé depuis ma nomination en 1999.

Publications	Thème
Revues internationales :	
– JASIST (Hubert et Mothe, 2009)	Combinaison des recherches par concepts et mots-clés
Conférences internationales :	
– CLEF'10 (Cabanac *et al.*, 2010b)	Évaluation des SRI
– ECDL'10 (Palacio *et al.*, 2010b)	RI géographique et évaluation
– HSI'09 (Chevalier et Hubert, 2009)	Réutilisation d'expériences de recherche
– ICEIS'07 (Hubert et Mothe, 2007a)	Sélection du système efficace pour un contexte
– WEBIST'07 (Hubert *et al.*, 2007c)	
– ICSOFT'07 (Hubert et Mothe, 2007b)	Réutilisation d'expériences de recherche
– ECIR'02 (Benammar *et al.*, 2002b)	Gestion de profils d'interrogation
– HCI'01 (Hubert *et al.*, 2001)	Interface graphique pour la fouille de texte
Chapitres d'ouvrages internationaux :	
– AKDM (Hubert et Teste, 2010)	Manipulation multidimensionnelle
Workshops internationaux :	
– INEX'06 (Hubert, 2007)	RI XML
– INEX'05 (Hubert, 2006)	
– INEX'04 (Hubert, 2005)	
– INEX'03 (Pinel-Sauvagnat *et al.*, 2003)	
– TREC'02 (Benammar *et al.*, 2002a)	RI Web
Revues nationales :	
– Document numérique (Cabanac *et al.*, 2011b)	Évaluation des SRI
– Document numérique (Cabanac *et al.*, 2011a)	RI géographique et évaluation
– I^3 (Hubert et Mothe, 2006)	Identification des contextes pour lesquels un système est efficace
– I^3 (Chrisment *et al.*, 2006b)	Recherche sémantique à base d'ontologies
– ISDM (Benammar *et al.*, 2003)	Gestion de profils d'interrogation
Conférences nationales :	
– VSST'10 (Dudognon *et al.*, 2010b)	Recherche sémantique à base d'onotologies
– CORIA'10 (Cabanac *et al.*, 2010a)	Évaluation des SRI
– INFORSID'10 (Palacio *et al.*, 2010a)	RI géographique et évaluation
– CORIA'09 (Hubert *et al.*, 2009a)	Recherche sémantique à base d'ontologies
– EGC'09 (Hubert et Teste, 2009)	Manipulation multidimensionnelle
– VSST'09 (Hubert *et al.*, 2009b)	Recherche sémantique à base d'ontologies

Publications	Thème
Conférences nationales (suite) :	
– VSST'07 (Hubert *et al.*, 2007a)	Sélection du système efficace pour un contexte
– CIDE'07 (Hubert *et al.*, 2007b)	Combinaison de systèmes
– CORIA'06 (Englmeier *et al.*, 2006)	Identification des contextes pour lesquels un système est efficace
– CORIA'05 (Hubert *et al.*, 2005)	RI XML
– VSST'04 (Dkaki *et al.*, 2004)	Recherche de la nouveauté
– BDA'03 (Augé *et al.*, 2003)	Catégorisation automatique de documents
– VSST'01 (Benammar *et al.*, 2001)	Gestion de profils d'interrogation
– VSST'01 (Augé *et al.*, 2001)	Catégorisation automatique de documents
Chapitres d'ouvrages nationaux :	
– Information & visualisation : enjeux, recherches et applications (Bonnel *et al.*, 2008)	Identification des contextes pour lesquels un système est efficace
Ateliers nationaux :	
– EGC'05 (Chevalier et Hubert, 2005)	Identification des contextes pour lesquels un système est efficace
Autres publications (liées aux projets) :	
– DynamO Livrable 5.2 : Spécifications des outils d'annotation dynamique et de recherche dans les textes (Dudognon *et al.*, 2010a)	
– DynamO Livrable 5.1 : Spécifications des outils d'annotation dynamique et de recherche dans les textes (Hernandez *et al.*, 2009)	
– IRAIA Deliverable : D.3.1.7. System Specification (Mothe *et al.*, 2002a)	
– IRAIA Deliverable : D.3.1.8. (Pre-prototyping) (Mothe *et al.*, 2002b)	

3 Encadrement

Cette section présente les encadrements de thèses et de masters recherche auxquels j'ai participé depuis ma nomination en 1999.

- Thèses de doctorat :
  - Bachelin Ralalason (2006 – 2010), « Représentation multi-facettes des documents pour leur accès sémantique ». Cette thèse s'est décomposée en deux parties. Dans un premier temps, Bachelin Ralalason s'est intéressé à un modèle de documents pour leur accès dans le cadre d'usagers malvoyants. J'ai pris en charge son encadrement pour la seconde partie de sa thèse dans laquelle il a développé un modèle d'indexation de documents à la fois sémantique et dynamique.
    - Publications en lien avec les travaux présentés dans ce mémoire : 2 conférences nationales (VSST'10 (Dudognon *et al.*, 2010b) et CORIA'09 (Hubert *et al.*, 2009a)), 3 rapports de contrat (Projet Dynamo) et 1 rapport de recherche.
  - Anis Benammar (1999 – 2003), « Profils en recherche d'information : définition, exploitation et adaptation ». Cette thèse a conduit au développement d'un modèle de profil d'utilisateurs dans le cadre de systèmes de recherche d'information.
    - Publications en lien avec les travaux présentés dans ce mémoire : 1 revue nationale (ISDM (Benammar *et al.*, 2003)), 2 conférences internationales (ECIR'02 (Benammar *et al.*, 2002b) et HCI'01 (Hubert *et al.*, 2001)), 1 workshop international (TREC'02 (Benammar *et al.*, 2002a)), 1 conférence nationale (VSST'01 (Benammar *et al.*, 2001)) et 1 rapport de contrat (IRAIA D3.1.8 (Mothe *et al.*, 2002b)).

- Masters recherche :
  - Mohammad El Makssoud (2010), « Partage et Capitalisation d'Expérience pour la Recherche d'Information ». Ce stage s'est intéressé à la modélisation des expériences de recherche des utilisateurs et à sa mise en œuvre à l'aide du système de gestion de bases de données Oracle.
  - Moussa Baraou Hamza (2007), « Mécanismes de bouclage de pertinence dans un système de recherché d'information XML ». Ce stage s'est inscrit dans le domaine de la RI XML. Il a consisté en une étude de solutions permettant de prendre en compte des jugements de pertinence formulés par l'utilisateur sur les résultats d'une première recherche pour améliorer l'efficacité d'une nouvelle recherche.
  - Randa Al Sabbagh (2005), « Gestion de versions de profils dans un système de recherche d'information ». Ce stage a visé à proposer une représentation des expériences de recherche des utilisateurs en s'appuyant sur les modèles de gestion de versions d'objets complexes.

Liste des figures

Liste des tableaux